



---

Statistical Policy  
Working Paper 14

**Workshop  
on Statistical Uses  
of Microcomputers in  
Federal Agencies**

---

Prepared by  
Subcommittee on Statistical Uses of  
Microcomputers in Federal Agencies

Statistical Policy Office  
Office of Information and Regulatory Affairs  
Office of Management and Budget

April 1987

**MEMBERS OF THE FEDERAL COMMITTEE ON  
STATISTICAL METHODOLOGY**

**(June 1986)**

**Maria E. Gonzalez (Chair)  
Office of Management and Budget**

**Barbara A. Bailar  
Bureau of the Census**

**William E. Kibler  
National Agricultural  
Statistics Service**

**Yvonne M. Bishop  
Energy Information  
Administration**

**David Pierce  
Federal Reserve Board**

**Edwin J. Coleman  
Bureau of Economic Analysis**

**Thomas Plewes  
Bureau of Labor Statistics**

**John E. Cremeans  
Office of Business Analysis**

**Jane Ross  
Social Security Administration**

**Zahava D. Doering  
Defense Manpower Data Center**

**Wesley L. Schaible  
Bureau of Labor Statistics**

**Daniel E. Garnick  
Bureau of Economic Analysis**

**Fritz Scheuren  
Internal Revenue Service**

**Terry Ireland  
National Security Agency**

**Monroe G. Sirken  
National Center for Health  
Statistics**

**Charles D. Jones  
Bureau of the Census**

**Thomas G. Staples  
Social Security Administration**

**Daniel Kasprzyk  
Bureau of the Census**

**Robert D. Tortora  
National Agricultural  
Statistics Service**

## PREFACE

The Federal Committee on Statistical Methodology was organized by OMB in 1975 to investigate methodological issues in Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in their personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues and that are open to any Federal employee who wishes to participate in the studies. Working papers are prepared by the subcommittee members and reflect only their individual and collective views.

The Subcommittee on Statistical Uses of Microcomputers in Federal Agencies organized a one-day workshop held on April 24, 1985. This working paper is based on the workshop and discusses four topics: planning to buy and use microcomputers for statistical purposes; electronic data dissemination; applications of microcomputers; and expert systems. The report is intended to provide helpful guidance to Federal agencies in purchasing and using microcomputers for statistical purposes.

The Subcommittee on Statistical Uses of Microcomputers in Federal Agencies was chaired by Terry Ireland of the National Security Agency, Department of Defense.

**MEMBERS OF THE SUBCOMMITTEE ON  
USES OF MICROCOMPUTERS IN FEDERAL AGENCIES**

Terry Ireland\*, Chair  
National Security Agency

Ken Berkman  
Bureau of Economic Analysis

Michael Leszcz  
Internal Revenue Service

Jay Casselberry  
Energy Information Administration

Tom Nagle  
Internal Revenue Service

Frederick J. Cavanaugh  
Bureau of the Census

Ronald Steele  
National Agricultural Statistics  
Service

Lawrence H. Cox  
Bureau of the Census

Peter Stevens  
Bureau of Labor Statistics

Richard Engels  
Bureau of the Census

Linda Bouchard Taylor  
Internal Revenue Service

Maria E. Gonzalez\* (ex officio)  
Office of Management and Budget

Mark Winer  
Office of Management and Budget

\*Member, Federal Committee on Statistical Methodology

## ACKNOWLEDGMENTS

The idea of a workshop as a focal point for proceedings on Statistical Uses of Microcomputers was suggested by Maria Gonzalez, Chairperson of the Federal Committee on Statistical Methodology. She also provided contacts in many Federal agencies, which made possible a broad Federal participation in the workshop.

The planning of the workshop was done by the Subcommittee. Four topics were selected for the sessions of the workshop. The chairpersons designated by the Subcommittee organized each session. They were:

### *Chairperson*

Session on Planning

Lawrence Cox,  
Bureau of the Census

Session on Electronic Data  
Dissemination

Ken Berkman,  
Bureau of Economic Analysis

Session on Applications

Ronald Steele,  
National Agricultural  
Statistics Service

Session on Expert Systems

Terry Ireland,  
National Security Agency

The proceedings were prepared by the chairpersons and rapporteurs of each session based on input from the speakers. The Subcommittee thanks all the speakers in the workshop for their participation.

Terry Ireland, who chaired the Subcommittee, and Norman Glick edited the final report.

Linda Taylor ably handled all the organizational and administrative details of the workshop -- the real basis for a very smooth-running conference.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

WORKSHOP ON STATISTICAL USES OF MICROCOMPUTERS  
IN FEDERAL AGENCIES

April 24, 1985

TABLE OF CONTENTS

	Page
Preface . . . . .	1
Members of the Subcommittee on Statistical Uses of Microcomputers. . . . .	ii
Acknowledgments . . . . .	iii
Introduction. MARIA E. GONZALEZ, Office of Management and Budget . . . . .	1
Session on Planning . . . . .	3
Summary. Prepared by FREDERICK J. CAVANAUGH, Bureau of the Census. . . . .	3
Introduction. LAWRENCE H. COX, Bureau of the Census. . . . .	5
The Census Bureau Microcomputer Information Center. RONALD SWANK, Bureau of the Census . . . . .	6
The National Security Agency Personal Computing Information Center. KATHY SCHNAUBELT, National Security Agency . . . . .	11
Use of Microcomputer Technology at the Bureau of Labor Statistics. PETER STEVENS, Bureau of Labor Statistics. . . . .	13
Discussion. LAWRENCE H. COX, Bureau of the Census. . . . .	23
Questions and Answers . . . . .	25
Session on Electronic Data Dissemination. . . . .	29
Summary. Prepared by JAY CASSELBERRY, Energy Information Agency. . . . .	29
Use of Microcomputer Disks to Disseminate Information. STUART WEISMAN, National Technical Information Service . . . . .	29
Cendata: Development and Implementation. BARBARA ALDRICH, Bureau of the Census . . . . .	34
Electronic Dissemination of Perishable Information. ROXANNE WILLIAMS, U.S. Department of Agriculture . . . . .	38
Questions and Answers . . . . .	40
Session on Applications . . . . .	45
Summary. Prepared by THOMAS NAGLE, Internal Revenue Service . . . . .	45

Spreadsheets and Statistical/Econometric Applications in Econometric Research. LINDA P. ATKINSON, U.S. Department of Agriculture . . . . .	46
Spreadsheets and Data Base Applications Used by the Crop Reporting Board in Reviewing Survey Indications and Preparing Publications. GARY NELSON, U.S. Department of Agriculture. . . . .	50
Manager's Perspective on the Acquisition and Use of Microcomputer-Based Graphics Packages. RICHARD W. HAYS, Internal Revenue Service . . . . .	51
Current Applications of UNIX-Based Microcomputer Systems. BRIAN CARNEY, U.S. Department of Agriculture . . . . .	54
Equipped for the Future? PAUL DOBBINS, U.S. Department of the Treasury. . . . .	56
Concerns About Data Integrity, Security, and Accessibility in an Environment Where Microcomputers and Mainframes Are Interfaced. DICK SHIVELY, U.S. Department of Agriculture. . . . .	58
Questions and Answers . . . . .	61
Session on Expert Systems . . . . .	67
Summary. Prepared by NORMAN GLICK, National Security Agency . . . . .	67
Introduction. TERRY IRELAND, National Security Agency. . . . .	69
Expert System Tutorial. GEORGE LAWTON, Army Research Institute. . . . .	70
An Extension of Statistical Software to Expert Systems. JAMES J. FILLIBEN, National Bureau of Standards . . . . .	78
Editing and Imputation. BRIAN GREENBERG, Bureau of the Census . . . . .	85
Discussion. MARK WINER, Office of Management and Budget. . . . .	93
Questions and Answers . . . . .	94
Appendix. Announcement of Workshop on Statistical Uses of Microcomputers in Federal Agencies. . . . .	97



## INTRODUCTION

**Maria E. Gonzalez, Office of Management and Budget**

A subcommittee of the Federal Committee on Statistical Methodology organized a one-day workshop on statistical uses of microcomputers in federal agencies. The purpose of the workshop was to share information among federal agencies on the statistical uses of microcomputers.

About 200 persons from federal agencies attended the workshop. The audience had an opportunity to ask questions and make comments in the discussion period of each session. All were acquainted with the uses of microcomputers. Some were also responsible for the planning of statistical uses of microcomputers in their agencies. The announcement of the workshop is included in the Appendix.

Four topics were discussed at this workshop.

1. *Planning of Statistical Uses of Microcomputers.* The first session described three microcomputer information centers in federal agencies. The purpose of personal computer (PC) information centers is to familiarize the agency users with the PC potentialities. This session focused on planning, implementation, and evaluation within federal agencies of statistical uses of microcomputers. The main questions asked were: Who should have microcomputers? For what purposes should microcomputers be used? In what configurations? At what costs? How will microcomputers coexist with central automatic data processing services?

2. *Electronic Data Dissemination.* This session dealt with different data dissemination methods. The discussion covered each agency's approach to data dissemination and the problems encountered in implementation.

3. *Applications of Microcomputers.* This panel discussion focused on the usefulness and weaknesses of microcomputer software and operating systems, the interface of mainframes and microcomputers, and factors affecting data integrity, security, and accessibility.

4. *Expert Systems.* The methodological basis for expert systems was discussed and several examples were given. The examples describe current expert systems with statistical applications.

The proceedings of this one-day workshop follow. For each session there is a summary, the presentations, and the discussions that followed.



## SESSION ON PLANNING

### SESSION SUMMARY\*

The microcomputer technology of the 1980s is a personal and, therefore, a user-oriented technology. However, planning for microcomputer technology is often very complex and causes many changes in the workplace. Program planners must take many factors into account when planning the introduction of a microcomputer system into their organization. Three personal computer information centers were described:

The Census Microcomputer Information Center of the Bureau of the Census

The Personal Computer Information Center of the National Security Agency

The microcomputer system of the Bureau of Labor Statistics

The planning, management and evaluation of microcomputer technology at the Census Bureau officially began in 1983 with a meeting of the Executive Staff. Prior to that time, microcomputer technology testing and evaluation work was ongoing at the Census Bureau, but this was the first time that agency-wide distribution of microcomputers was discussed. The Census Microcomputer Information Center (CMIC) was established as a result of this meeting. To give greater emphasis to the importance of microcomputer technology, the Census Bureau located the Center in the Office of the Director with its manager reporting directly to the Associate Director for Administration.

The purposes of CMIC are to assist employees in learning about microcomputer technology -- both from a user point of view and a manager/procurer point of view -- and to reduce the overall costs of microcomputer technology purchase and maintenance. Employees are given access to various brands of hardware and software to test prior to purchasing. They are also given "hands-on" experience in the use of the newest in microcomputer hardware and software through special arrangements made with the various vendors and manufacturers. On-site training in the use of hardware and software is provided by outside trainers, with the divisions paying the costs for their employees. Costs currently range from about \$100 to \$125 per person per day, which are quite favorable in comparison with commercial costs of similar training.

The activities of the National Security Agency's Personal Computer Information Center (PCIC) started approximately 18 months ago, when NSA

---

\* Frederick J. Cavanaugh, Bureau of the Census.

established the PCIC to train employees in the use of PCs and vendor-developed software. It did not take long to discover problems of compatibility among various brands of microcomputers. Therefore, standards were established to ensure that:

1. All microcomputer systems at NSA are compatible with one another for effective communications and portability.
2. All systems are able to function using the UNIX operating system -- again, to allow for communications and portability.
3. The microcomputer systems are supportable; that is, they must be easily and cheaply repaired.
4. The systems are secure, so as not to divulge secret information.

NSA has set its microcomputer standards around the IBM PC and PC/XT in a UNIX-based environment (IBM's PC/IX) and its office automation standards around the Wang PC.

BLS's microcomputer system is essential for efficient office operation, and BLS has kept this in mind in designing and developing its system. The BLS Executive Staff is very supportive of the microcomputer system.

In designing the microcomputer system at BLS, several critical needs have to be met. These include:

1. The need for a system that can readily provide terminal communication with mainframe computers.
2. The need for a system capable of communicating among various machines and those located in field offices as well.
3. The need to provide security for confidential information.

BLS undertook research and experiments to determine which microcomputer system best met its needs. Upon completion of the research, a single system comprised of machines from a single manufacturer was implemented and a set of standards was developed around its operation and use. The present system includes over 100 IBM PC/XTs and three Ethernet (FIPS 107) local area networks.

The microcomputer systems described in the presentations form a continuum from the *experimental or user-oriented* approach to the more standard *production or program-oriented* approach. However, despite a commonality of needs and objectives, each agency has chosen a different approach to planning and managing microcomputer technology.

## INTRODUCTION

Lawrence H. Cox, Bureau of the Census

Welcome to the Workshop on Statistical Uses of Microcomputers in Federal Agencies, sponsored by the Federal Committee on Statistical Methodology. We begin with this session on *planning*.

Microcomputer technology is the technology of the 1980's. It is a *personal* and, therefore, a *user-oriented* technology. However, its focus on the individual often can be misleading from a planning perspective -- at the agency or office level, planning and managing the use of microcomputer technology becomes very complex very fast. While encompassing important technical issues concerned with hardware, software and communications networks, this technology also quickly brings the planner face-to-face with the business of managing and deriving improvements systematically from technological change. Inevitably, the introduction of microcomputers into an organization changes the workplace and the skills and orientation of workers. It presents new choices and often demands that these be made swiftly.

In large organizations and small offices, the following questions must be addressed:

- where does microcomputer technology fit into the agency or office?
- how should it be introduced?
- how can the organization experiment and grow with this technology?
- what must the organization do to plan and manage this technology effectively?
- should standards be set for its use? which standards? how should they be set? by whom? how should they be enforced?
- what sort of future decisions need to be made, and who should make them?

We are fortunate today to have a panel of experts in this field, whose experience should shed light on answers to these and other important questions facing the statistical program manager about to embark on the introduction of microcomputers into his or her organization.

They speak with the experience of individuals tasked with managing groups assigned these responsibilities in three different Federal agencies: the Bureau of the Census, the National Security Agency, and the Bureau of Labor Statistics.

The speakers are:

- Mr. Ronald Swank, Manager, Census Microcomputer Information Center, Bureau of the Census.
- Ms. Kathy Schnaubelt, Chief, Information Resources and New Technology Branch, National Security Agency.

and

- Mr. Peter Stevens, Chief, Division of Communications and Computing Technology, Bureau of Labor Statistics.

Until recently Kathy was Chief of NSA's Personal Computer Information Center and had direct responsibility for the functions we will discuss this morning. Ron and Peter have had these responsibilities on a continuing basis for some time.

Each speaker will make a brief presentation on how the problem of planning the use of microcomputers was addressed in their agency. I will follow with a few comments by way of formal discussion, and we will then open the floor for discussion and questions from the audience.

**THE CENSUS BUREAU MICROCOMPUTER INFORMATION CENTER**  
**Ronald Swank, Bureau of the Census**

The words "microcomputer" and "personal computer" are often used in a manner that blurs their intended use. In the true sense of the word "microcomputer," the Bureau of the Census has been using microcomputers since 1968. The FOSDIC (Film Optical Sensing Device for Input to Computers) allowed us to film and input census forms to computers without manual data entry. In 1973 we attached IBM 6250 tape drives to Sperry mainframes, approximately three years before Sperry announced similar availability, and experimented with the ATL automated tape library. In 1982, eight Apple II+ personal computers were used to do the Puerto Rican Economic and Agriculture Census data checking and editing. These projects established the feasibility of using microcomputers in much of the Bureau's work.

The Bureau's organization (3500 employees at headquarters, 9000 nationwide) can best be described as 35 separate companies (divisions, in our parlance) sharing the same resources, computers, management services, etc. You can imagine the problem this presents in setting priorities, standards and general directives. All of the Census Bureau's funding does not come directly from Congressional appropriations. So there has been a great deal of discussion on the best way of introducing microcomputer technology to the Census user community, funding it and not intimidating or alienating Bureau users. In 1983, a joint decision was made to establish the Census Microcomputer Information Center (CMIC). The Center with a staff of 4 was placed organizationally in the Director's Office for two reasons: (1) to show Executive Staff support for technology and encourage users to make

active efforts to become familiar with its capabilities and (2) to avoid turf battles.

The CMIC is a clearinghouse of information for use by Census Bureau employees.

The goals and objectives of the CMIC are to:

- assist Bureau employees in their analysis of microcomputers;
- provide access to and demonstrations of a variety of hardware, software and peripherals;
- provide hands-on experience with microcomputers without capital investment by the individual divisions;
- provide training on microcomputer hardware and software;
- provide a clearinghouse for documentation, catalogs, and pointers to knowledge for microcomputers, end-user computing and office automation;
- decrease the cost of hardware and software through more informed procurement decisions.

With the direction of program managers, Census Bureau employees may visit the Center for information about microcomputers, for discussions of the characteristics of particular computers and the applicability of microcomputers to projects, or for hands-on experience on a variety of machines in an attempt to implement those projects. One can use the computers in the Center for weeks if necessary, experimenting with various software on different machines. The role of the Center is to help Census Bureau staff define their processing needs, advise them of applicable software and guide them towards suitable computer equipment. The CMIC contains the more popular microcomputers and the more popular software. Yet, there are significant numbers of microcomputers that may provide a unique perspective in the industry and may offer the best overall systems for a particular problem. Therefore, the CMIC also sponsors product demonstrations about those microcomputers that are not currently on display in the Center.

#### **CENTER OPERATION/USE**

The Center's hours of operation are 9:00 a.m. to 4:00 p.m. Census Bureau personnel may schedule time to use a particular machine, software package, tutorial or specialized peripheral device for one-hour segments. They may also request one of the Center support personnel to work with them. We generally have personnel in the Center from 7:00 a.m. to 5:30 p.m. Time before and after hours of operation is devoted to Center personnel, allowing us to gather and exchange information on the day's occurrences and to provide specialized support to executives.

Some of the typical questions arising on a given day may be as simple as:

What's the difference between a hard disk and a floppy?

How can I get specific information on a specific product and its capabilities?

What kind of tutorials/training are available for Lotus, dBase, etc.?

Why doesn't a package perform in a specific manner?

#### WHAT'S AVAILABLE IN THE CENTER

The Center subscribes to approximately 40 periodicals dealing primarily with microcomputers and associated technology. About 20% of these magazines are provided free. Also the Center has a library of 300+ books dealing with microcomputers, hardware, software, peripheral devices, etc. These books are directed at all levels of personnel. The magazines and books are available for checkout by Census Bureau employees.

The Center subscribes to Data Pro for microcomputer hardware and software. There are numerous other vendor- or industry-provided catalogues available for review in the Center:

- IBM Personal Computer and XT Software Guide
- The Blue Book for IBM
- Engineering and Scientific Progress
- The Book of Apple Software
- The Ratings Newsletter
- IBM Software Directory

Many of the supply and peripheral device catalogues are provided by vendors. Public domain software is available in the Center. Most of it was acquired from Capital-PC for IBM's and compatibles and the Freeloader 500 software for the Apple machine. This software is not copyrighted and is available for the cost of reproduction. We have found many useful utilities available that have saved our users much development time.

Microcomputer software in the following categories is available:

Communications software	Mathematical
Database management systems	Specialized
Electronic spreadsheets	Statistical
Integrated software	Word processing
Presentation graphics	Utilities
Programming languages	

This software is available for user evaluation. The end user determines whether the product will produce the required results. About 15% of our software was provided by vendors for use in the Center -- but only in the Center -- for evaluation purposes and not for production work.

The Bureau's policy on copyrighted software is that it is not to be copied for any reason other than backup.

## **HARDWARE**

There are 2 IBM PC/XT's hooked to a local area network. A Sperry Model 50, a Wang PC, a Grid, Apple Macintosh, peripheral devices, plotters, printers, Polaroid palette, etc. are available.

Many microcomputer vendors (43 to date) have come to the Census Bureau to demonstrate their products, and many have loaned their products for evaluation from 30 to 60 days, depending on product. Some of the vendors are:

A & F Computers	Sony
Digital Equipment	Fujitsu
Olivetti	Motorola
Hewlett-Packard	Exxon
Data General	Radio Shack

## **ELECTRONIC BULLETIN BOARD**

In February 1985, an electronic bulletin board was placed into service to facilitate information interchange on product evaluations, user projects, etc.

## **PROCUREMENT POLICY**

The Census Bureau's procurement policy evolved because of our organizational structure and our funding. While all procurement actions are to be processed and controlled through the Bureau's Procurement Office, requests for ADP-related actions will continue to require some specialized processing.

The justification and acquisition approval for microcomputer equipment and off-the-shelf software and supplies totaling less than \$10,000 is delegated to the Associate Director level. The ADP staff no longer is required to review and approve such purchases. When the purchase order is sent to the Procurement Office, requests for sole source and brandname purchases costing more than \$500 must include a brief justification. When the purchase order is received in the Procurement Office, information copies are forwarded to the Census Microcomputer Information Center to be used to update the Census Bureau inventory of microcomputer equipment and software.

## **MICROCOMPUTER MAINTENANCE POLICY**

The Census Bureau's microcomputer maintenance policy is based on cost. For every six machines purchased we purchase a spare machine because the cost of a one-year maintenance contract on the first six equals the cost of the spare. These machines are not just stored; they are used in noncritical environments where they can be removed to replace a critical machine as needed within one hour. When a user encounters an equipment problem that is beyond the user's capability to resolve, he or she contacts our Technical Services Division (TSD) service representative who will respond by sending a technician to the user's site to isolate the cause of the equipment problem.

If it is something simple that the technician can repair on the spot (such as replacing a fuse, reseating a loose board, or tightening a plug), the technician will make the repair. If the problem cannot be resolved by the technician on site, the technician will telephone the CMIC to request that a replacement computer or input/output device be loaned to the user until the user's machine is repaired. TSD will set up the replacement equipment for the user (if necessary) and take away the machine that needs repair. The user should be able to resume normal operations with minimal delay, aggravation and frustration.

If the device is still covered by its original warranty, TSD will arrange to have it repaired under the terms of the guarantee. If the warranty is no longer valid, TSD will arrange to take the machine to a designated dealer for a repair estimate. When the machine is left with the dealer, a hand receipt will be signed by the dealer and returned to TSD. When the dealer calls the estimate to TSD, TSD will prepare a purchase request and forward it to the user's division. The division will insert the appropriate accounting code, approve the action, place a priority flag on it, and send it to the Procurement Office. The Procurement Office will expedite all micro maintenance requests by calling the dealer with a purchase order number. When the repairs have been finished and the machine is ready for pickup, a driver will take the purchase order to the dealer and pick up the machine. This procedure is valid for any repairs totaling less than \$1,000. In cases involving repair estimates in excess of \$1,000, TSD will contact the microcomputer user to discuss whether the repairs should be authorized and, if so, what procedure must be followed.

The loaner machine will be under the control of the CMIC with the following priorities governing their use:

Top priority -- to any user where TSD has removed a machine for authorized repairs.

Second priority -- for use in support of hands-on training classes sponsored by CMIC.

Third priority -- for use by someone who wants to do small projects on a borrowed machine.

Priority will mean exactly that. A broken machine will be replaced with a loaner from the CMIC even if it means having to take the loaner away from someone who is using it under a lower priority. I want to emphasize that this is our current policy, but it can be changed very quickly. We are constantly monitoring this procedure and continually reassessing our options (i.e., outside service contract).

**MICROCOMPUTER TRAINING SUPPORT**

We established a classroom with 16 machines for hands-on-training. We did this because of the numbers of people requiring training and the cost of sending people to outside courses. The types of courses taught are: Introduction to Microcomputers, Databases, Word processing, Spreadsheets, Graphics, etc.

Originally there were requests for training of 3000 persons in all aspects of microcomputers. That has been reduced to approximately 2100. We believe this training demand will be high initially and then will drop off dramatically. Outside instructors have been hired to teach our classes. We have had a great deal of success with this process because of the quality of instructors acquired. To pay for this training facility we charge back directly to the user division the cost of the instructor, software purchased and maintenance cost of the classroom. This cost goes to a maximum of \$125 per class, significantly cheaper than to send all people to outside training.

NOTE: Many vendors sell at a small cost educational licensing agreements providing copies of their software for each machine in the classroom. Some vendors will not do this; then we must purchase copies for each machine at full price.

#### **OFFICE AUTOMATION**

I have specifically not addressed the topic of office automation, as we are still planning and discussing exactly what office automation is going to mean at the Census Bureau. Our primary planning focus at this time is to determine what functions need to be provided Bureau-wide and what functions will be left to individual operating units.

#### **THE NATIONAL SECURITY AGENCY PERSONAL COMPUTING INFORMATION CENTER Kathy Schnaubelt, National Security Agency**

The National Security Agency established a Personal Computing Information Center (or PCIC for short) approximately a year and a half ago. This action was taken in response to the Agency's growing demand for personal computer products.

In the year prior to the opening of the PCIC, many new personal-computer products and vendors were reaching the marketplace. A growing number of these products were in turn being purchased by a cross-section of Agency elements. This mix of products across the Agency began surfacing problems such as that of system incompatibility. This may be illustrated by the example of a diskette of data or software running on one computer brand but ;not on a different brand of computer. The PCIC was designed to assist Agency personnel in the selection, acquisition and use of an established set of "standard" personal computer products.

The basis for the selection of standard products was determined by the Agency's needs as a whole. One such requirement was for the UNIX operating system. Hardware selected as the Agency standard workstation would have to be able to run under the UNIX operating system. At the root of decisions of this nature was the concept of compatible hardware and software products that would be easy for people to acquire.

Another important concern for us was security. By going to standardization, that problem may be minimized by the selection of products that meet this requirement and then training personnel to use them.

A third consideration was supportability. Maintaining a variety of microcomputers, or personal computers, can be a logistics nightmare; stocking of parts, replacing them, etc., in any number can be devastating.

Finally, there is cost. By limiting the number of kinds of personal computers and software products that we use, we are able to buy large numbers of each at a lower per-unit cost. Right now we have thousands of microcomputers in the Agency, and we have plans to buy many more, which should result in a significant savings from bulk buys.

The PCIC was established to meet the following objectives: 1) to promote the use of standard equipment; 2) to share and centralize our small systems resources (like everyone here, we have a limited number of people to support these products); 3) to minimize the end-user application load; 4) to maximize cost effectiveness; and 5) to centralize product registration (providing anonymity in our workplace).

The PCIC has become a focal point for all Agency standard products, and to date these products include: an Agency standard terminal/workstation which is an enhanced IBM XT; the standard office automation equipment which is the WANG Professional Computer; an interim standard local area network. So there will be a family of Agency standard host computers.

The PCIC provides its customers with information on all of the standard products that are available; and this includes a reference collection of books, periodicals, in-house-developed working aids, research guides, comparison charts of the capabilities of the different products, and a referral service for technical questions. It also provides demonstrations of standard products. Anyone can go down to the PCIC and use one of the standard products, whether it's hardware or software.

To encourage the use of the PCIC by Agency personnel, the PCIC tries to make the acquisition of standard commercial products as simple as possible. Rather than have each office go out and do their own purchase request, an authorized individual can come into the PCIC and request commercial software. The software is actually stocked in the PCIC. We have licensed some items (like CONDOR and MICROPRO products for example). By doing that, we have actually reduced some costs by 70%.

Non-standard products may still be purchased, but on a limited basis. A non-standard product must be requested in writing. This request is reviewed by a software evaluation team to determine the validity of the purchase request. When a product offers a unique capability, it is purchased and evaluated. A favorable evaluation results in the product's being added to the list of standard products. A product which does not offer any capabilities beyond the standard product line, or in fact is defective, would be placed on a prohibited-purchase list. In any case, the PCIC still does the actual purchasing, whether it's for a standard product or an evaluation copy of a non-standard product. This saves the requester from the paperwork of writing a purchase request document.

While the purpose of the PCIC is to furnish standard products, it also functions in identifying products that meet certain minimum requirements for Agency use. These products are added to the list of standard products to provide a flexible work environment for Agency personnel. The goal is not to restrict what people do or how they do it, but to make sure that the products they use are compatible with other products used throughout the Agency.

**USE OF MICROCOMPUTER TECHNOLOGY  
AT THE BUREAU OF LABOR STATISTICS  
Peter Stevens, Bureau of Labor Statistics**

I made the discovery when putting this talk together that I could take the various displays and shuffle them and present them in almost any order I chose. I'm not quite sure what the conclusion from that would be, but with this heady sense of freedom, I decided to start in the middle. Therefore, the first display you see discusses a brief introduction as to where we are now. The Bureau of Labor Statistics has approximately 100 microcomputers, almost all of them standard IBM PC/XT's (see Display 1). We also have three Ethernet Local Area Networks, two in D.C. and the other in the San Francisco regional office. We have network licenses and centralized software libraries for all of these machines. This is one point, and the first of the points which I will be emphasizing, where some of the things that we are doing that are, perhaps, different from what is commonly done. Floppy disks have no essential role in the entire operation. If I had my way, I wouldn't have them.

**Bureau of Labor Statistics  
Networks and Microcomputers**

**Where we are now**

**Approximately 100 microcomputers in use (mostly highly modified IBM PC/XT's)**

**Three Ethernet (FIPS 107) Local Area Networks, two in DC, the other in San Francisco. Software libraries are centralized.**

**We are close to completing the "large scale pilot" stage of our development effort.**

**For each application area our goal is to identify and validate quality products which can be made a part of the standard BLS microcomputing environment.**

*Display 1.*

In general, the way people get software onto their machines is through the local area networks from centralized storage devices. We are getting now to the end of what might be called the "research phase" of this entire new technology operation. The three networks were all acquired by a competitive

procurement which we ran a couple of years ago and which is, in effect, a large-scale test.

That gets to the last point on Display 1, which is the basic goal for what we are trying to accomplish right now: to identify and validate quality products which can be made a part of the standard BLS microcomputing environment; then, in the next stage of our operations, to make standards for use throughout the Bureau.

When I looked at Display 2, I decided I could put it up and talk about it for twenty minutes without any trouble at all because it enumerates the applications and I think that gives some scope of the project. But given the terrible time constraints that we are under, I will spare you a lot of discussion here.

The following are major application areas:

- Word Processing
- Graphics
- Spreadsheets
- Statistical Analysis
- Data Base Management
- Survey Data Collection
- Survey Control
- Project Management
- Calendar Management
- Network Services, including Electronic Mail, Shared Data Management and Inter-network Routing.
- National Communications via Public Value-Added Networks (X.25 & FIPS 100 standards).
- Mainframe Communications Gateways for Interactive and Batch Operations.
- Access to the Local Networks from remote (usually portable) microcomputers.

#### *Display 2.*

However, there are two things worth pointing out. Some may know from the previous references that "FIPS" stands for Federal Information Processing Standards, which are produced by NBS and which we are trying to follow. We have more standards than FIPS 100, and those things are, in general, a significant part of our operation.

One other point, before moving on here, that I think is worth some mention: applications like word processing, graphics, and spreadsheets are standard and well known; but the applications that I call here Survey Control, Project Management, and Calendar Management get into a function for the microcomputer which I don't think has gotten the emphasis it deserves. This is a Control and Management function. In the same sense that a microcomputer is a useful tool to use with a project management package, it is also used and useful for keeping track of one's personal calendar and the

ordinary flow of activities through the division. As we find users responding to technology, this is definitely a growing area.

Anyway, enough for the present. The reason for Display 3 is not so much a chance to give you the details of how the Bureau operates, but to make a point that our efforts in these areas were started in response to a serious and well-understood operational problem that we are having. The large, centralized mainframe computer provides, in our view, a very poor, very weak environment for the general area of interactive applications.

### **How This All Got Started**

Throughout the 70's the Bureau's approach to computing relied upon two large, IBM-mainframe, computer centers accessed via dial-up telephone lines.

While this environment served the large-scale, batch-oriented, survey processing well, other applications were served poorly:

Interactive applications were very hard to develop, and response from the mainframe computers varied widely.

Data communications were a constant source of problems, especially those with our Regional Offices.

The proliferation of incompatible word processing equipment caused continuing operational problems and prevented any more ambitious office automation efforts.

The most promising technical approach to solving these problems was:

Powerful microcomputers for interactive processing.

Local Area Networks for the heaviest communications and for configuration management.

Internetwork and Mainframe Gateways for extended communications.

Public Data Networks for national communications.

### *Display 3.*

Again, I'm sure you wouldn't like to see me stand here and cry, so I'll spare you the details of the problems we have had with data communications since the AT&T divestiture.

The final point under the problem areas is again worth some emphasis. We have, I think, some thirteen odd different brands of word processors in place. None of them communicate with each other. This is a story that has

been, again, welltold. There was, in the Bureau's top management and operations management, a perception that this had caused us a great deal of difficulty and a very strong desire not to perpetuate that same sort of incompatibility and lack of communication in the new technology.

The lower part of Display 3 shows briefly what we have selected as the technological underpinnings of the steps we are taking. Again, we could have a long discussion on say, minicomputers versus microcomputers and the local area network services, but it is beyond the scope of this panel. I will only mention that these issues were very seriously considered, and the choices listed were not made lightly.

I would like to draw your attention to the phrase "configuration management." Having, let's say, several hundred microcomputers all using the same software packages would not be, in our view, sufficient to guarantee compatibility.

Companies are constantly issuing new versions, and these new versions are frequently incompatible with each other. So you need not only to standardize with the level of machinery, but you need to do version control and configuration management to insure that the potential of a standard environment endures. One of the major functions of the local area network is that it makes it really possible to do this. If we wish to put up a new version of a particular procedure, we can do so. We can test it and then make that transition very easily.

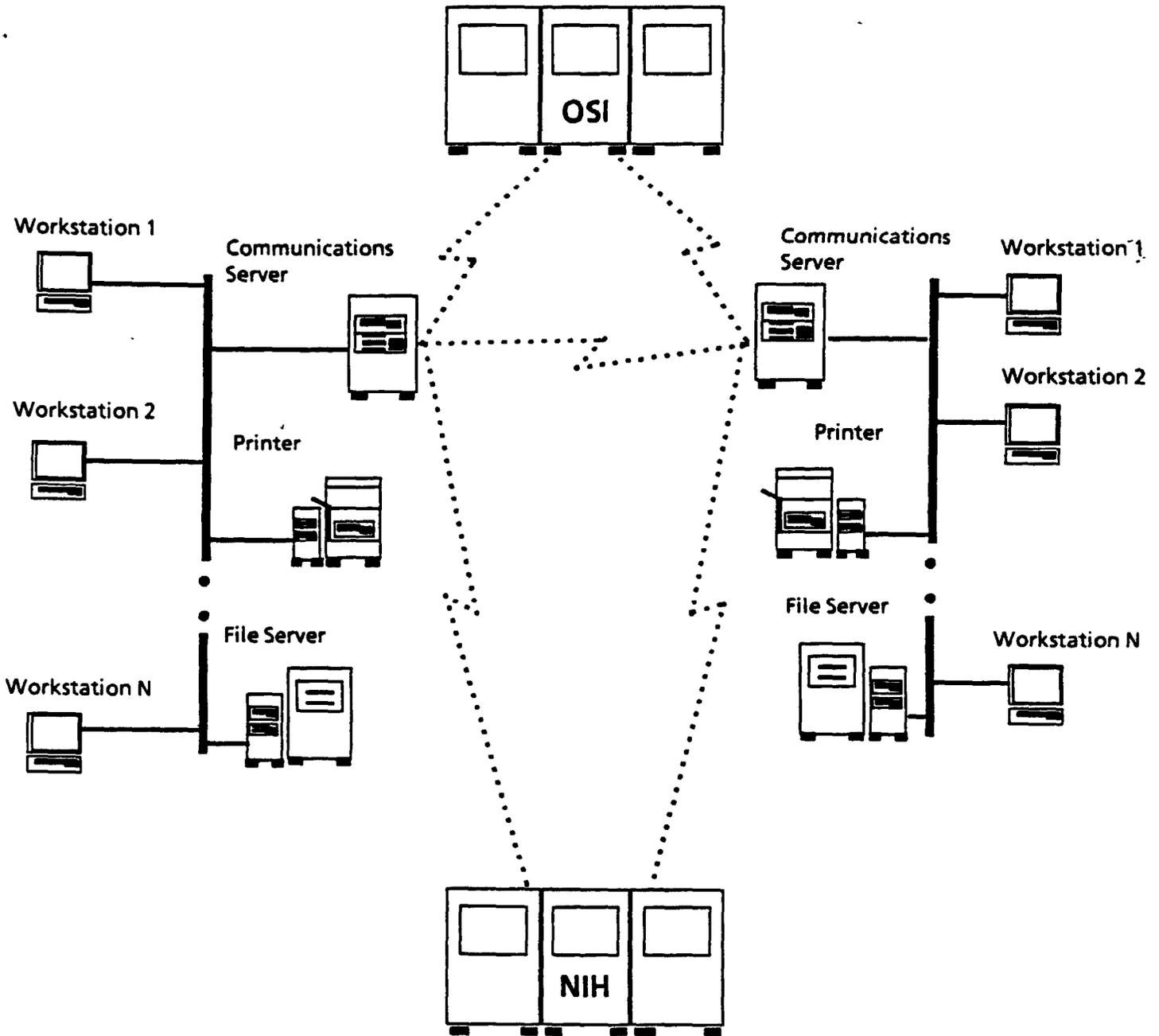
Back when I was planning this, I had visions of myself running down the hall with 500 floppy discs trying to distribute them. It was the horror of that nightmare that led us in that direction.

Display 4, "How This All Got Started" is from a configurations perspective. I urge you not to take this too literally, but, in conjunction with Display 3, it does demonstrate the basic structure of the communications and technical environment. The large, vertical black bars indicate the local area networks themselves (that is, cable connections between machines in a single area). We use two computer centers: National Institute of Health and Optimum Systems, Inc. Those dotted lines indicate communications through the public telephone system.

On the networks themselves we basically have two types of devices: The workstations (that is, machines that people use) and network services for file storage, printing, communications, etc.

Now we are at the point where we can get down to the most important part of this presentation. One of the things that I would like to try and share with you, from our experience, is an idea that I call, on Display 5, "Important Operating Assumptions." An assumption here means about the same thing that "theory" means in physics or chemistry. It means an idea that we believe and accept as true and act upon, but at the same time are constantly retesting and reevaluating.

# INITIAL LOCAL AREA NETWORK CONFIGURATION



Display 4.

## Important Operating Assumptions

No single supplier can come even close to supplying top-quality products for all our requirements.

The best quality and most creative software development now is being done by Independent (and frequently quite small) Software Vendors.

Standards, de facto and formal, play a much more important role for the microcomputer market than they do for the mini or mainframe market.

We can increase effectiveness and reduce risk by emphasizing open systems and standards rather than by becoming locked in to one manufacturer's product line.

The most reliable source of information about new products is our own testing.

The selection, testing and integration of hardware and software are professionally very demanding tasks. Statisticians and economists should not have to become microcomputer experts to use the equipment well.

Quality in the initial selection of hardware and software is only the start of an effective operation. Support, maintenance, and especially release control for software are essential to long-term effectiveness.

Planned and controlled redundancy is the best and, in many cases, the only way to achieve high reliability.

### *Display 5.*

The first four items are a basic description of why we are interested in "open systems" or open-systems interconnection. We have substantial experience with being in the tender and enveloping grasp of a single manufacturer and in discovering that that manufacturer's products don't meet new needs, or that there is no way to interface some new piece of equipment to the existing equipment.

**THE MOST RELIABLE SOURCE OF INFORMATION ABOUT ANY PRODUCT IS OUR OWN TESTING.** This point belongs in bold print because that is probably the essence of the whole project.

The computer business has always been full of what I will call "hype": statements of doubtful truths made just to sell equipment. The microcomputer business is, if anything, worse than the mainframe side of the business. We have found that things like articles and advertisements in magazines, the flowing promises of salesmen, and similar frivolities are simply not a basis upon which we can operate. We have certain responsibilities to our users in the Bureau so that when we say something is going to work, they can expect that it will work. We can't then turn and

say that the salesman said it will work. Much of our validation is this testing of the product claims.

The next two items on Display 5 deal with another very important aspect of our work. Doing the kind of validation that will cut through the hype is, in our view, a demanding task and not one which need be or should be placed upon the working statistician and economist. We have a very large number of users that want to use this technology. We have a much smaller number that wish to become microcomputer experts. We are trying to create an environment in which economists, statisticians, managers, clerical personnel, and the whole BLS community can use microcomputers effectively without having to go through the struggle and pain that is associated with selection, testing, and integration of the underlying technology.

The last item in Display 5, I think, is very similar to the ones already expressed by Census. The way you get the reliability is through redundancy. One of the conclusions that followed from that idea is to use a standard configuration. Even though a particular machine may be intended for word processing and the machine next to it may be intended for statistical analysis, the underlying hardware will be the same. So that, if on the day the analysis is due, that particular machine decides to go out to lunch, the other machine can be used to finish the job.

We are getting down toward the end, so we can summarize this by talking about the Project Goals and Current Policies (Display 6). You may remember that I mentioned there were three important problems that this research effort was attempting to address: the need to have an environment in which we could create good interactive systems; the need to deal with our data communication flows; and a need to provide effective intercommunication between machines when used for statistical survey work, office automation, or any other purpose. Those were the goals and the motivation to start the project. They remain the goals. Every product we distribute must be thoroughly tested before full regional use. Some of the regional offices have very little background in data processing. What we put there had better work, because we don't have the travel budget to fix the mess if it doesn't.

## Project Goals and Current Policies

### Project Goals:

To solve the identified major problems with communications and interactive computing.

To ensure that new products are thoroughly tested before being put into production systems or into all Regional Offices.

To open up new application areas, especially in the areas of end-user computing and office automation.

To establish the basis for the continuing, orderly introduction of improved hardware and software.

### Current Policies:

The selection, evaluation, procurement, and support of new products is centralized. Strong, de facto standards exist.

The development of end-user applications is decentralized.

The introduction of new products to Bureau production systems is closely managed. Pilot tests are required and high-level approval must be gained before production commitments are made.

The emphasis on compatibility, full communications, and Bureau-wide usage is quite strong.

### *Display 6.*

Finally, we see this whole technology as having opened up the potential to get into kinds of applications, that simply weren't being done at all by any type of computer, such as some of those personal and local organizational ones that I mentioned earlier. We now need to establish a basis so that we can continue to introduce, in an effective and orderly manner, new products and new technology that continue to pour out of the industry.

From that, we have certain policies: the centralized selection, evaluation, procurement, and support of new products. There is some doubt as to whether we will be able to sustain a centralized procurement function because of some of the problems in government procurement which are beyond the scope of this presentation. In contrast to this centralization, the development of end-user applications is decentralized. That is, the way that persons use the machines for a particular personal or organizational task is a matter of their judgment and their discretion.

When we are talking about introducing this technology into Bureau production statistical systems, there is much stronger management control; and developments are closely watched. We insist on Bureau testing and evaluation before committing important Bureau projects to the new technology.

I think I have said enough about the need for compatibility.

Finally, on Display 7, under the heading of Where We Are Going, there is basically more of the same. I mentioned we are getting toward the end of the large-scale research phase. We are planning to add local area networks into all eight regional offices instead of just San Francisco.

We have one aspect of the Bureau which may be unique in that the Commissioner of Labor Statistics has a PC in her office. She also has one at home and uses them both. She has an intense personal interest in what I call here, "Management Communications." Through the local networks we have possibilities that we never had before.

Through the research phase of this work, we have not had what I might call "traditional government procurement cost/benefit justification analysis" very much. I expect, as we move to the broader expansion of microcomputers into Bureau activities, that analyses of that nature will become important. There are many areas about procurement issues that are, at the moment, looking through a glass very darkly.

#### **Where We Are Going:**

**As the performance of specific hardware and software products is validated, their use will be expanded to production tasks.**

**The number of Local Area Networks will be expanded to include all Regional Offices.**

**The communication facilities will be expanded to include Cooperating State Agencies for data collection and survey processing.**

**Management communications, among the Commissioner, Office Chiefs and Division Chiefs, will become increasingly important.**

**The number of microcomputer workstations will be significantly expanded. Obsolete or ineffective equipment will be replaced by microcomputers.**

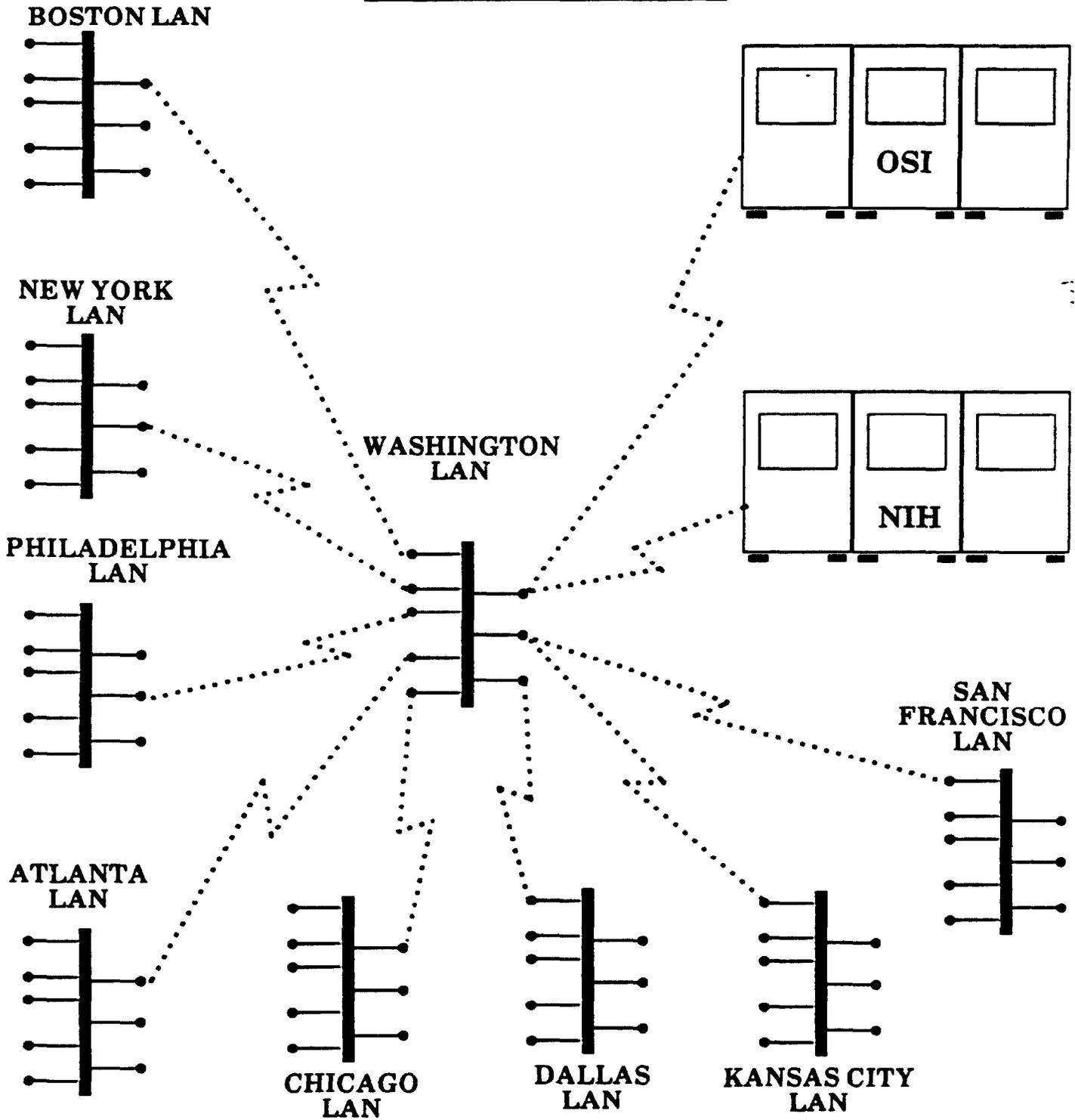
**New hardware and software developments will be watched for possible replacements to standard products.**

**As the new technology replaces existing equipment and applications, greater emphasis will be placed on cost/benefit justifications.**

#### *Display 7.*

Display 8 shows where we expect to go technologically. I ask you not to take that too literally. This is not a technical model, but rather a demonstration of the way we see things getting done with each of the regions having its own network communicating to our network in Washington.

# FINAL NETWORK CONFIGURATION



Display 8.

## DISCUSSION

Lawrence H. Cox, Bureau of the Census

I will attempt to keep my comments brief so that we can have a full interchange between the speakers and the audience in proper "workshop" fashion. In proper "discussant" fashion, I will highlight what I see as the major similarities and differences among the three approaches taken, in the context of what I have learned from the presentations collectively and from my experiences at the Census Bureau.

I have learned that microcomputer technology is a *must* for statistical programs. Automated, interlinked statistical program offices are more efficient and effective than those which are not. Users of statistical information have discovered microcomputer technology; and, so, statistical data providers have a responsibility to keep pace. Data review and analysis at its best is an interactive process between the expert data analyst and the data, supported by statistical software. Mainframe computing cannot offer these services on a large scale in a realistic manner or at a competitive price.

I have learned that an organizational locus is needed to provide information and support both to management and users as this new technology becomes introduced and assimilated within the organization. We have seen that such a group can have any of several functions, depending upon organizational size, needs, goals and objectives:

- user education and handholding
- repository of literature
- source of hands-on experience
- maintenance
- training
- develop and distribute product lists and recommendations
- establish guidelines for microcomputer procurement, use, maintenance, training, etc.
- recommend standards for microcomputer hardware, software and uses of microcomputer technology
- establish and enforce such standards
- aid in the procurement process
- evaluate procurement requests
- decide upon procurement requests
- advise in the management of this new technology
- play an active role in its management

These functions, as I have presented them, lie on a continuum from the more passive, permissive or *experimental* approach to the more standardized, structured, or *production-oriented* approach. These needs and the management philosophies underlying them seem to me to be well-represented on that continuum by the three agencies represented here today.

The *free-market* or *laboratory* approach adopted by the Census Bureau says, in effect, let's provide our diverse group of programs and users with the information necessary to begin to explore uses of microcomputer technology. Let's minimize the procurement obstacles to doing so, and let's work closely

with users in their applications and see what lessons are to be learned and what patterns emerge. In effect, as an organization, let's not force microcomputer hardware and software choices, but let's closely manage and monitor several experiments and learn from each of them.

At the National Security Agency, decisions were driven by the overriding need to standardize on hardware and software choices sufficiently to allow diverse and distant groups to talk to each other and access the same data and programs, but stopped short of imposing *inessential* standards. Within a predefined architecture of standards, NSA users are free to experiment, to share information and to tailor choices to programmatic and individual needs.

At the Bureau of Labor Statistics, the requirements for good and standard communications between offices and geographic areas were paramount. Experiments were conducted to fix upon the best choices, from which standards are to emerge. The environment is intended to be uniform and capable of supporting continuing, production-oriented work.

Reflecting upon this continuum for a moment, I could equally describe it as being from *user-oriented* to *program-oriented*, reflecting a progression defined in terms of the number of diverse programs and functions within these agencies which each agency seeks to address with automation at the microcomputer level.

Interesting, all three organizations share several characteristics: they are not small, they deal routinely with massive amounts of data, their paramount concern is improved and broader access to their own data, their systems require mainframe gateways or links, and they operate under strict data security requirements. However, for reasons which we have heard and others you may explore in open discussion, they have chosen three different approaches to tackling the problem of planning and managing microcomputer technology.

#### QUESTIONS AND ANSWERS

**Q1: How was the Census Bureau able to acquire 500 microcomputers in a little over one year given GSA guidelines?**

A1 (Mr. Swank): The Census Bureau did not go around GSA guidelines and standards, but worked within the existing regulations. Most procurements are off the GSA schedule.

**Q2: What variety does the Census Bureau have in their brands of microcomputers?**

A2 (Mr. Swank): Currently there are 25 different brands of microcomputers in operation at the Census Bureau.

**Q3: Does Census go through the "GSA microcomputer store" in procuring its microcomputers?**

A3 (Mr. Swank): Yes, when possible. However, the GSA microcomputer store does not stock all brands, and this forces the Census Bureau to go elsewhere.

**Q4: Why did Census create a separate staff for microcomputers when they already had an established automatic data processing staff?**

A4 (Mr. Swank): The Executive Staff of the Census Bureau wanted to show support for microcomputer technology and to give it high visibility and, therefore, created the Census Microcomputer Information Center and placed it in the Director's Office.

**Q5: Has the Information Center taken an active role in education of upper-level management in the uses of microcomputer technology?**

A5 (Mr. Swank): Yes, each member of the Executive Staff has been given at least an introductory course on microcomputer usage.

**Q6: The presentation left several unanswered questions that should be addressed:**

1. What about the lack of a management system for electronic files?
2. How are archiving and disposition of files handled?
3. What about programming for the PC's?

A6 (Mr. Swank): Electronic filing systems will come in the near future. There are several such systems in existence now, but the costs are astronomical.

A6 (Mr. Stevens): Software for record retention currently exists, but the big problem is file retention for which very little software is available.

**Q7: Are the PC's at Census "stand-alones" or are they networked?**

A7 (Mr. Swank): Some PC's are networked others are "hardwired" to the mainframe; the majority are "stand-alones."

**Q8: Two questions regarding the presentations:**

1. What is meant by "software standards"?
2. Some software packages need improvements, corrections, etc. In each agency, does anyone speak to the manufacturers as a representative of the agency?

A8 (Mr. Stevens): "Software standards" means software standards. For example, there are at least three subcategories of word processing software, and each would have a separate software standard at BLS.

A8 (Mr. Swank): Corporate licensing would be the answer. Those manufacturers that will not discuss corporate licensing have so much business they do not need to help and keep the client happy.

A8 (Ms. Schnaubelt): The focal point for NSA is with the vendor rather than the manufacturer. NSA has had problems with RUBIX from IBM. The smaller vendors are much more eager to get the business and give better contractual terms than the large firms.

**Q9: Is there a very strong recommendation from the panel for a PC information center?**

A9 (Dr. Cox): An independent PC information center is an absolute necessity in a large organization.

A9 (Mr. Swank): Each agency definitely needs at least a resource person if not a center.

**Q10: Would a small group need a PC information center?**

A10 (Dr. Cox): Not necessarily a center, but at least a reference person.

**Q11: Regarding machine-oriented versus people-oriented use of microcomputers, what would the individual agencies do for the people? What are the goals?**

A11 (Mr. Swank): At the Census Bureau, if the individual divisions have the budget, they will get the microcomputers they ordered within 30 days of the request.

A11 (Ms. Schnaubelt): The goal is to have a PC on each desk.

A11 (Mr. Stevens): At BLS, the only drawbacks to a microcomputer on every desk are budget and procurement.

**Q12: With the advent of work-at-home, is there a use of portable PC's for this purpose?**

A12 (Dr. Cox): The major problem with portable PC's for take-home use is data security -- a large problem for each of the agencies represented.

A12 (Ms. Schnaubelt): At NSA, portable microcomputers are used by executives and others, but these machines are kept "clean" (i.e., they have never had any sensitive data on them). The portables are used for training purposes only.

A12 (Mr. Swank): The Census Bureau has many "checkout" machines, but some of these are secure machines and cannot be taken out of the building.

A12 (Mr. Stevens): BLS definitely believes in the work-at-home concept and has machines for this purpose. However, precautions are taken to protect confidential data.

**Q13: How are services provided to field operators?**

A13 (Mr. Stevens): The regions do their own training on the uses of the BLS system.

A13 (Mr. Swank): There is a standardized configuration of microcomputer technology in each regional office with a nationwide company contracted to carry out maintenance.

A13 (Ms. Schnaubelt): Data and software are transmitted world-wide by mail or other secured means of communication.



## SESSION ON ELECTRONIC DATA DISSEMINATION

### SESSION SUMMARY\*

The second session dealt with electronic data dissemination, focusing on disseminating information for use with microcomputers. While the first panel discussion focused on how agencies use microcomputers within their own internal environments, this session deals with the impact of microcomputers on users of federal agencies' data and the possibilities for agencies to make information available for microcomputer users (that is, dissemination of data using floppy discs or through telecommunications).

There are some very interesting opportunities for federal statistical agencies to use new media to provide data to users more quickly and in a form that is more highly usable than current printed methods. The three speakers will deal with these issues. The first speaker is from the National Technical Information Service (NTIS) which is primarily an archival-type agency for disseminating federal data and information. The NTIS program to disseminate data on floppy discs, the problems encountered, and the various issues surrounding this area will be discussed.

The second speaker is with the Bureau of the Census and works with their telecommunications system called CENDATA. CENDATA is used to distribute perishable Census information to users.

Our final speaker is from the Department of Agriculture. She will describe the current, ongoing process to implement a contract with the Martin Marietta Corporation to establish a telecommunications system for the dissemination of large databases containing agricultural information.

### USE OF MICROCOMPUTER DISKS TO DISSEMINATE INFORMATION Stuart Weisman, National Technical Information Service

The history of the National Technical Information Service (NTIS) dates back to 1945 with the establishment of a publication board to assist in making unclassified government documents available to the private sector. The program went through various transformations, reaching its current status as an agency of the Department of Commerce in 1970.

---

\*Jay Casselberry, Energy Information Agency

The law creating NTIS states that NTIS is to search for, collect, classify, coordinate, integrate, record, catalog, and disseminate information. In the early 1970's, NTIS received its first machine-readable information product. In 1981 a new unit was established within NTIS to manage its product line of data base files and software. In the summer of 1984 NTIS began to sell data on floppy discs.

The current NTIS machine-readable-products program contains about 10 bibliographic data bases, 300 source-text non-bibliographic data bases, 800 numeric and statistical data bases, and 1300 computer software programs. With this substantial amount of information available, NTIS began a review of procedures for disseminating information products for microcomputers.

The following criteria were considered when NTIS reviewed the potential for disseminating their information products on microcomputer diskettes:

- o Forecasts of the number of microcomputers
- o Forecasts of the primary type(s) of microcomputers being used by business and professionals
- o Physical size of the computer diskette
- o Microcomputer operating systems
- o In-house and/or contractor production of diskettes
- o Information products to be made available on diskettes
- o Entire and/or subsets of information files made available
- o Production of microcomputer software
- o Whether to reformat the data for use with popular data base spreadsheet formats

NTIS has decided to make information products available on 5 1/4 inch diskettes for IBM and IBM-compatible microcomputers. Diskettes are produced by a contractor, and costs are determined based on the number of diskettes required.

The main problems that have been encountered are in the loss or incorrect conversion of data when tapes or diskettes are produced, mishandling of diskettes during shipment, and improper use of the diskettes by customers. The way to overcome these problems is to establish procedures for checking a diskette against the original magnetic computer tape, and to instruct transportation companies and end-users on the proper handling of diskettes.

In the future NTIS will consider producing information products on high density diskettes, hard discs, and, where it is practical, optical or video discs.

With the future increases in microcomputers by business and professionals, NTIS is making a long-term commitment to having information products available for microcomputer users. With the proliferation of data

management and analysis being done with microcomputers, NTIS recognizes the needs of this user community. Displays 9 through 16 illustrate the work of NTIS.

#### HISTORY OF MACHINE-READABLE INFORMATION PRODUCTS

Late 60's First machine-readable products arrive at NTIS

Early 70's Production Group formed to process orders for machine-readable products

Late 70's Concept of Product Management introduced

1981 Office of Data Base Services

1983 Video disc products from NASA

1984 Data files available on diskette

*Display 9.*

#### DATA TAPES

Over 1,000 Titles	32 Source Agencies
40 Titles Updated Annually	25 Titles Updated 2-6 Times a Year
15 Titles Updated Monthly	Remainder Updated Less than Annually
Standing Orders Available	

*Display 10.*

#### MAJOR DATA COLLECTIONS

National Center for Health Statistics (NCHS)

Federal Communications Commission (FCC)

Energy Information Administration (EIA)/ U.S. Department of Energy

National Bureau of Standards (NBS)

Human Nutrition Information Service/ U.S. Department of Agriculture

Defense Logistics Supply Center/ U.S. Department of Defense

Federal Reserve Board (FRB)

Environmental Protection Agency (EPA)

*Display 11.*

**DECISIONS, DECISIONS, DECISIONS!!**

Size: 5 1/4" vs. 8 1/2" (3 1/2" not readily available)

Density: Double vs. single sided;

Single vs. double density (quad-density not readily available)

MS-DOS vs. CP/M (or MS-DOS vs. PC-DOS)

Total in-house vs. contracting-out vs. in-house/out-house balance

Products pre-selected vs. demand-driven selections

Complete files only or subsets/extracts

Software

ASCII only or various DBMS/spreadsheet/formats

*Display 12.*

**DATA DISKETTES**

5 1/4" Diskettes

Standard ASCII Format

For IBM-PC Microcomputer

Unique Accession Numbers Assigned

Data Tapes Converted to Diskettes

Documentation Required

*Display 13.*

**PLAYER RESPONSIBILITIES**

<b>NTIS</b>	<b>Contractor</b>	<b>Source Agency</b>
Order Input & Control	Create diskette master	Provide master tape or diskettes (with appropriate documentation)
Copy tape to be used for conversion	Archive Master	Available for consultation
Ship Orders (with documentation)	Duplicate Master	
Available for consultation	Get duplicates to NTIS	
	Available for consultation	

*Display 14.*

The Action

Customer contacts NTIS -- "Available on Diskette?"

YES	NO
1. Price	1. Estimate price (based on # of diskettes)
2. Customer orders	2. Customer orders
3. Order to contractor	3. Copy master tape
4. Contractor duplicates master	4. Order to contractor with tape
5. Duplicate to NTIS	5. Contractor creates master diskette and duplicates master for customer order
6. NTIS mails (with documentation) to customer--overnight delivery	6. Duplicate to NTIS (price is actual # of diskettes)
	7. NTIS mails (with documentation) to customer--overnight delivery

Display 15.

Problems

Original tape	----->	Bad tape from agency
Copy tape at NTIS	----->	NTIS error in copying tape
Contractor converts tape to diskette master and duplicates master	----->	Contractor error in conversion process or duplication process
Duplicated diskettes sent to NTIS	----->	Problems created in handling of diskettes
NTIS ships diskettes to customer	----->	(magnetic field, dropped, smudge, coffee, etc.)
Customer receives and processes diskettes	----->	Customer mishandles diskettes (see above) plus diskette processing

Display 16.

**CENDATA: DEVELOPMENT AND IMPLEMENTATION**  
**Barbara Aldrich, Bureau of the Census**

CENDATA is an information system for disseminating Bureau of the Census ("Census") information electronically. Development of CENDATA began in mid-1983 when Census decided that certain data, especially time-sensitive economic data, should be available on-line. CENDATA was developed under the guidelines that the data should be available on-line as soon as possible after release and that the system developed should be done at no cost to Census.

The system was proposed as non-sole source (i.e., not limited to only one contractor). In addition, no money was to be involved in the arrangement with any contractor, and Census was to have control over the information made available. During the entire process of developing the specifications and establishing memoranda of understanding with qualified vendors, Department of Commerce lawyers assisted in refining the language and procedures.

Census' list of qualifications for vendors wishing to access CENDATA and make the information available included:

- o A CENDATA user should only have to pay for time used accessing CENDATA
- o CENDATA should be available separate from other data bases, be clearly identified, and include the entire CENDATA package
- o CENDATA must be available seven days a week
- o A CENDATA vendor must be willing to accept data delivery via telecommunications
- o A CENDATA vendor must be able to offer its users the services of national telecommunications networks
- o The ~~system~~ system must be an end-use-based, user-friendly system

The reasons behind the above qualifications were to:

- o ensure that vendors did not add hidden fees or package CENDATA with other services
- o enable users to use major telecommunications networks to minimize costs
- o obtain vendors with the capabilities to handle a large-scale data base such as CENDATA
- o increase dissemination of Census information products.

Of the dozen vendors who have shown interest in the CENDATA system, four met the criteria established; and memoranda of understanding have been signed with two.

The first vendor, Dialog Information Services, went on-line with CENDATA on August 1, 1984. (Dialog is extremely prominent in the library community.) Dialog has CENDATA available using the standard menu-based system and also makes the information available in a full-text-searchable format.

In mid-October, 1984, the Glimpse Corporation made CENDATA available. Glimpse, in cooperation with the Chemical Bank of New York, markets data to the financial community.

With the success achieved by the first two vendors in expanding the dissemination of Census data, Census is anticipating adding new vendors who service different sectors of the public. With the inherent advantages of CENDATA over traditional publications, Census hopes to continue to expand its user network.

The primary advantages of CENDATA are the timeliness of the data and the ease of using the system. One of the first goals of CENDATA was to have sensitive economic information available within minutes after any embargo on the information is lifted. Examples of the type of sensitive information available are manufacturers' and shippers' orders, retail sales, housing starts, and balance of payments.

Having this information available electronically assists users who are located away from Washington where the information is initially disseminated in press releases. The data are available weeks before users would receive it in published form, and it can be downloaded into a user's standard information system for review and analysis.

Census also maintains an inventory of its products on CENDATA. This allows a user to quickly determine if a particular publication has been released, and, if so, the price, source, and Government Printing Office stock number.

The illustrations that follow, Displays 17 through 21, show how CENDATA has been developed for ease of use. Menus are designed to provide an inexperienced user with a choice of selections, and to move from general to the more specific. In addition, instructions are provided to help a user move through the system.

#### **THE CENDATA INTERACTIVE SYSTEM**

**The Online Information Utility at the U.S. Census Bureau.**

**A very small portion of the Census Bureau's vast data holdings has been included in this "information utility."**

**Do you wish to see the CENDATA menu? If yes, enter Y or (return). If not, enter LOGOFF to end session.  
?Y**

*Display 17.*

-- CENDATA MAIN MENUS

- 1 Introduction to Census Bureau  
Products and Services
- 2 What's New in CENDATA
- 3 U.S. Statistics at a Glance
- 4 Press Releases
- 5 Census User News
- 6 Product Information
- 7 CENDATA User Feedback
- 8 General Data
- 9 Agriculture Data
- 10 Business Data
- 11 Construction and Housing Data
- 12 Foreign Trade Data
- 13 Governments Data
- 14 International Data
- 15 Manufacturing Data
- 16 Population Data

Enter item number or ? for help.  
?15

*Display 18.*

15--MANUFACTURING

- 1 Introduction to the  
Manufacturing Statistics  
Program
- 2 M3 Preliminary Report, July 1984  
.  
.  
.
- 8 Aluminum Ingot and Mill Products,  
June 1984 (CIR M33-2)

Enter item number or ? for help.  
?2

*Display 19.*

15.2--M3 PRELIMINARY REPORT,  
JULY 1984

- 1 M3 Narrative Summary
- 2 Value of Manufacturers' Shipments
- 3 Value of Manufacturers New Orders
- .
- .
- 7 Ratio of Manufacturers' Inventories  
and Unfilled Orders to Shipments

Enter item number or ? for help.  
?3

Display 20.

15.2.3--August 30, 1984  
TABLE 2, PART 1: VALUE OF MANUFACTURERS'  
NEW ORDERS FOR INDUSTRY GROUPS, MARKET  
CATEGORIES, AND SUPPLEMENTARY SERIES

		--Seasonally adjusted-- Monthly (Millions of dollars)		
SIC Code	Industry	Jul. 1984 (p)	Jun. 1984 (r)	May 1984
	All manufacturing industries.	192,450	190,620	193,680
	<b>Manufacturing industries</b>			
	with unfilled orders .....	103,496	102,051	104,482
	Durable goods industries ....	100,489	99,171	102,256

--more--

Display 21.

After moving through the choices of information topics, the user is presented with the information requested.

An experienced user may move through CENDATA more quickly by specifying all parameters of its search at the same time. For example, by specifying 15.2.3 initially, all menus may be bypassed; and the user moves directly to manufacturing (15), the M-3 report (2), and specifically the value of manufacturers' new orders (3). This development allows CENDATA to provide the necessary information and instructions for novice users without unduly hindering more experienced users.

As with any developing system, Census is soliciting comments from actual and potential users to determine possible system improvements and expansion of the data base. The primary users at the current time are economists, industry analysts, and market researchers.

Future plans are to expand the data base with additional Census products. Upcoming products to be added are 1984 country population estimates and statistical profiles of every country in the world. With the addition of the statistical profiles, CENDATA moves into a new area since the information is from the International Data Base rather than from a publication, and the profiles are not readily available outside the system.

**ELECTRONIC DISSEMINATION OF PERISHABLE INFORMATION**  
**Roxanne Williams, Department of Agriculture**

The Department of Agriculture has as a primary function the dissemination of information about conditions related to agriculture. The Extension Service is one way the Department uses to get information disseminated at the local level. In addition, the Department has long utilized the printed media for the dissemination of information around the nation. A few years ago, a number of agencies in the Department became dissatisfied with the print media because of the difficulty in getting information to interested parties as quickly as necessary. The agencies, acting independently, tried electronic communication of data. Use was made of a number of commercial services such as DIALCOM, AGNET, and AGRADATA. DIALCOM is equivalent to an electronic bulletin board. AGNET is an on-line information system developed at the University of Nebraska.

About two years ago, the Department started to have problems with the use of these services. Other information companies wanted the Department to provide the data going to existing services. They did not want to have to go to competitors for the information for a variety of reasons. One reason was that they wanted to be able to say they obtained the data directly from the USDA. Supplying each potential vendor with USDA data was just too much of a burden for the Department.

In order to continue to get data to the ultimate end user and at the same time meet the needs of commercial vendors, it was decided to establish a single department-wide system of electronic data dissemination. No agency will be forced to use this system; but if an agency decides to use electronic media, it must use the Department's system. This central system will then service the commercial vendors, including DIALCOM, AGNET, and AGRADATA.

The Department decided to limit the scope of the project to what we call "time-sensitive perishable data." One example of this type of data is the agriculture marketing reports. These are perishable because they contain the current prices and the current sales of all the different commodities around the country. The data are in constant demand and they are constantly changing as new reports arrive continuously. The demand for the quick and timely dissemination of these data is very high.

The Department is utilizing a commercial vendor, Martin Marietta Corporation, to provide this service. This maintains a Department policy of not allowing public access to the Department's computer. It also keeps the Department from establishing a service that can be adequately provided by the private sector. Martin Marietta acts as an agent of the Department and has agreed not to use its position in order to benefit itself in the dissemination of these data to ultimate users. Martin Marietta can only disseminate these data through the system established for the Department. Other commercial vendors (we call them Level I users) can tie into the system with auto-dial or auto-set facilities. For a price, they can even have the main system's computer call their computer as soon as data are released and transfer those data immediately. Thus all vendors will have excellent and "equal" access to USDA's perishable data.

Equal access also meant to us that Martin Marietta would not charge other commercial vendors outrageous prices for access to the system. We wanted to keep the costs to Level I users reasonable. Martin Marietta was very reasonable and agreed to modest and uniform charges.

Ease of access was also important to the Department. In order to maintain simplicity and keep programming costs low, we decided to use a straightforward file structure for the data with access obtained through a menu-driven system. The resulting simplicity of the system not only makes for easy access by users, but it also allows originating offices within the Department to upload files with a minimum of effort.

Further, the originating offices maintain complete control over their own data in the system. They determine when data go into the system, when they are to be released, and when they are to be deleted. Martin Marietta only maintains the hardware and software of the system.

In addition to meeting the requirements of outside (Level I) users, the system has been designed to the Department's own internal requirements for information. A second type of user (Level II) has been defined. Level II users are primarily offices within the Department and the Extension Service. Other Federal agencies which make heavy use of agriculture data will be included. In order to service the Level II users, we asked Martin Marietta to allow access to smaller segments of data. These users do not need to obtain bulk data by telecommunications. The system allows us to break down bulk reports into smaller segments all of which are accessible via simple menus.

The Department anticipates that the effect of the new system will be manifold. Users should have much better access to a wider range of information. Internal communication of information within the Department should improve significantly. The demand for hard copy should be significantly reduced. All of these effects should help to reduce the cost to the Department of data dissemination.

## QUESTIONS AND ANSWERS

**Q1: What were the particular problems with mailing floppy discs; what kind of reject rates were encountered; and, if the discs are used for data transfer, how much of a backup do you need?**

A1 (Mr. Weisman): Some problems in handling of the discs during shipment may have been avoided because we chose to use an overnight delivery service instead of the Postal Service. The quality of the service has been very high, there is very little handling required, and the service has not failed yet.

**Q2: Did you mention that there were some bad discs that needed to be replaced?**

A2 (Mr. Weisman): Yes. It is very difficult to track down where the mishandling of discs actually occurred.

**Q3: Is there a flat percentage of reliability?**

A3 (Mr. Weisman): The percentage of problems is very small, but it does occur.

**Q4: Has NTIS considered direct phone transmission of data; that is, could users call directly to the NTIS computer, similar to commercial data bases?**

A4 (Mr. Weisman): We did make our bibliographic data bases available similar to what Census is now doing (as mentioned in the talk by Barbara Aldrich). That was started around 1974, or perhaps earlier. I believe there are now four vendors carrying our data base. In addition, NTIS encourages vendors to carry its statistical files and source files. To date, no vendor has elected to carry these files because it is more difficult to carry these files than a bibliographic data base. NTIS has no plans at this time to make these files available through telecommunications.

**Q5: What are the plans for disseminating data from the 1990 decennial census?**

A5 (Ms. Aldrich): In terms of data dissemination for 1990 decennial census data using CENDATA, there are no solid plans, but it is an issue for thought. The product information section of CENDATA could be used as a daily update or product release for 1990. I believe that there will be some electronic dissemination, but the amount and the level are not really being addressed at this time.

**Q6: Please tell us more about the software available through NTIS; is it public-domain software, software that the agencies have written for their own use, or some other type of software?**

A6 (Mr. Weisman): While I am the manager for data files and data bases and there is a separate product manager for software, I will try to answer your question. The criteria that NTIS uses for handling software are the same as those used for data files; that is, the software must be Government-produced. The software must also have a common usage and be useful to others.

**Q7: NTIS currently sells a catalog of public domain software for \$40 that includes quite a lot of information. Why doesn't NTIS publish separate catalogs of microcomputer software and mainframe software?**

**A7 (Mr. Weisman):** At the present time NTIS only has three packages available on diskettes for microcomputers, the rest are for mainframes. NTIS does not convert software at the present time and may never do so. Currently there are not enough diskettes available for microcomputers to justify a separate catalog.

**Q8: Does Census have any feedback from CENDATA users on the services and charges?**

**A8 (Ms. Aldrich):** Yes, based on discussions with users, the charges seem reasonable. DIALOG priced CENDATA at \$36 per hour, their most inexpensive commercial rate. That price does not include the telecommunications network charge which, with discount, is generally about \$6 per hour. The Chemical Bank version of CENDATA is priced at \$28 per hour and includes the telecommunications charge. In addition to the positive feedback we are receiving on prices, we receive feedback on what is in CENDATA, what users would like to see in CENDATA, and what they do not like.

**Q9: Is it possible to download CENDATA data and create other data files based on this?**

**A9 (Ms. Aldrich):** CENDATA is all public domain and no part is copyrighted. Therefore, it is available for users to download to their computers or add to other data bases. This caused a slight problem with DIALOG because so many of their data bases are copyrighted. To end any confusion, a notice was put in the DIALOG newsletter pointing out that CENDATA is in the public domain.

**Q10 (Mr. Berkman):** Would Barbara and Roxanne discuss the impact upon their particular agencies' personnel who generate the data, in transferring the data to the two systems they discussed?

**A10 (Ms. Aldrich):** I would like to cover the impact in two areas: the positives and the negatives. The negative for the people generating the data is that they must provide it to us in machine-readable form, either in the appropriate kind of floppy disc or via telecommunications to our microprocessor. There are some guidelines, with respect to designing tables that must be followed, which are quite difficult. The industry standard for CRT screens is 80 characters across, so any table must be defined in 75 characters since the vendors requested five characters for control. Often tables are split vertically, with the first part becoming Table 1, Part A; then the second part is Table 1, Part B; and so forth. The positive advantage to people preparing time-sensitive information and providing the data to CENDATA is a reduction in the interruptions from outside the agency with requests for data. Prior to CENDATA, when a data embargo was lifted, staff members would spend the remainder of the day answering the telephones and reading data over the phone. With the advent of CENDATA, users have an alternative where they can quickly receive the data. They can copy the data from CENDATA to their microcomputers and eliminate the need to listen to it over the phone and record it. There are both positives and negatives to the individuals who provide CENDATA with the information. In all cases the

individual division which is the source of the data provides the CENDATA staff with the information.

A10 (Ms. Williams): Agriculture has designed a system whereby each agency retains control over its own data. This is a very sensitive subject, so the system was designed so that each agency enters its own data into the system. Because of the wide variety of equipment used to process data and create reports by our agencies, the system also needed to be designed so that the agencies did not need to change their current methods of doing business. To accommodate the agencies, each agency only needs to put a header card on its report to identify the report. If a report is to be broken up into different levels of service, an additional header card is necessary. Based on the header card(s), the system knows how to handle the report that follows. One agency, the Agriculture Marketing Service, required another accommodation because it used a leased wire service with a special protocol. Current users of these data had taps on the wire which were usually linked to teletype machines. A microcomputer system was placed between their system and our system to convert the protocol and place the headers on the data. This allowed their system to operate exactly as it did prior to development of our system.

**Q11: Does CENDATA provide a computer tape to its vendors or is data communicated via telecommunications? Also, how often are the vendors' files updated?**

A11 (Ms Aldrich): All CENDATA are transmitted via telecommunications. We use an enhanced word processor with telecommunications capabilities. Information initially goes into a private file where it is integrated into our standard system. We review the system exactly as a user would see it and determine if there are any problems. Simple problems are corrected using the vendor's editor; serious problems may be corrected by deleting the file and starting over. When we give the go-ahead, the data become available on the vendors' systems. On DIALOG the files are brought up overnight so the data becomes available the next day. We update daily based on data to be made available and changes in our product listings. The update is controlled by a vendor's software. We move records into and out of their systems.

**Q12: Does the Bureau of the Census pay for the update costs?**

A12 (Ms. Aldrich): No. Census developed the menu. We work closely with the software design people at each vendor.

**Q13: Do the vendors limit the amount of information?**

A14 (Ms. Aldrich): Certainly not in the case of DIALOG. They have the philosophy that however much information you can give them they will accept it. They consider data storage to be cheap and pride themselves on being one of the largest vendors. In the case of Chemical Bank, they have not constrained us either. About once a year they request for planning purposes an estimate of how much storage we will need in the next two years. We have a small amount of data available on-line with a rich potential for it to get out of hand, but thus far there are no problems.

**Q14: What were the reasons Census decided not to go sole source?**

A14 (Ms. Aldrich): One of the primary reasons was our objective to get the system operational as quickly as possible. By offering it to several vendors, we could avoid the procurement process. Another appeal was that by going with several vendors, CENDATA would be available to different segments of the community. With different vendors it might be possible to reach users that previously had not been Census data users. I think that in the case of DIALOG we have found a lot of librarians who were not previously users.

**Q15: Has meeting the different protocol requirements of the different vendors involved much extra work?**

A15 (Ms. Aldrich): No, because we have only one system and one format for the data; each vendor must agree to adapt that format to whatever they see fit to use. There is one set of codes which are very simple and straightforward.



## SESSION ON APPLICATIONS

### SESSION SUMMARY\*

The relatively recent emergence of powerful microcomputers (micros) coupled with the availability of specialized vendor software packages for micros has significantly enhanced the federal statistical community's ability to gather, manipulate and analyze data. Today, more than ever, it has become easier to perform data analyses previously considered to be impractical due to resource and time limitations associated with traditional manual and computer methodologies. Accompanying enhanced analytical capabilities have improved methods for communicating the results of our data analyses. Powerful graphics software along with improved graphics plotters and color displays have made it possible to easily paint pictures reflecting data analyses, which before were only possible through relatively expensive and involved mainframe processing.

The boom in microcomputer usage in the areas of statistical and economic analyses is due in large part to the many advantages micros have over mini and mainframe computers. In particular, today's micros have storage capacities and processing speeds which often exceed mainframe capabilities commonly found just 10 years ago. Micros are generally simpler and easier to use than minis and mainframes; they are often portable; and they cost less to procure, operate and maintain. Micros are usually more reliable (less down time), and they often possess the ability to communicate with minis and mainframes, which permits micros to access and transfer large data files.

Along with the "hardware" advantages, there are also "software" advantages associated with micros. In particular, there is an abundance of high quality and user-friendly vendor software packages available, many of which permit the user to add his or her own code to modify and enhance the package's capabilities. Relative to mini and mainframe costs, these software packages are inexpensive.

A few disadvantages of micros should be mentioned as well. The ability to exercise security measures and ensure control appear to be more limited. Today's micros are slow in comparison to current state-of-the-art mainframes. There exist serious compatibility problems of file structures between vendor software packages. Finally, there is often an added personal cost to the micro user in the area of additional time spent in procurement and maintenance, since these activities are usually not required of a mainframe user.

The discussions which follow address many of the issues mentioned above.

---

\*Thomas Nagle, Internal Revenue Service

**SPREADSHEET AND STATISTICAL/ECONOMETRIC APPLICATIONS  
IN ECONOMETRIC RESEARCH**

**Linda P. Atkinson, U. S. Department of Agriculture**

Microcomputers are in widespread use throughout the Economic Research Service (ERS). I will be discussing their application not by secretarial staff for word processing or by data processing professionals, but rather by the economic research staff themselves.

Our economists first became involved with microcomputers through the use of spreadsheet software, and this is still where the bulk of the applications are. Packages such as Supercalc and Lotus 1-2-3 are used extensively for data preparation, developing tabular reports, producing high-quality charts, graphs, and plots, performing if-then analyses, and interfacing with mainframe software. Some of the systems which have been developed with these packages are, in fact, quite sophisticated.

One group, for example, has developed a program using Lotus 1-2-3 to assess preliminary economic impact of foreign pests to producers, consumers, and society in general. A partial budget analysis is used in which different economic scenarios are simulated by allowing changes in costs of production, yield, and prices for the affected crops. The entire system is menu driven and has options for various tables and graphs which can be produced. The program set-up is being used as a template from which similar analyses can be developed, such as a program to evaluate the impact of change in ozone concentrations on yields.

Another group had been using Supercalc for data entry and preparatory calculations before running a program on the microcomputer to convert the data to the form required for input to mainframe packages such as TROLL or SAS. After running these mainframe programs, files of output were then transmitted back to the microcomputer and reformulated for spreadsheet entry so that tables and graphs of output were automatically generated. Additional changes in the form of model output results could then be made, interfacing the flexibility of the microcomputer with the calculating power of the mainframe computer.

Now this group has a simplified version of their model, the world grain-oilseeds-livestock (GOL) trade model, running entirely on the micro in Supercalc. The GOL model is an annual simulation model consisting of 27 country and regional models and 20 major agricultural commodities. The individual models are linked to solve simultaneously for a vector of prices which clear world trade. The global model system has equations for 339 country-commodity combinations. Running a 20-year projection on the full linked model on an IBM PC/XT took 48 hours; however, an individual country model runs in about 15 minutes. They hope to improve speed considerably by the acquisition of an IBM PC/AT with memory upgrades. The program has been set up to ask questions of the user, such as what country is to be analyzed for what start and end dates. Users like the flexibility of the spreadsheet format; one can get in and look at a simulation, watch the numbers change and see where any problems are. Built-in equation writers allow you to change the structure of a model or you can edit it directly. You can pre-create graphs and have them contain historical data to compare to simulated results.

A good reference on building such models in spreadsheets is an article from the February 1985 issue of Byte magazine entitled "Simultaneous Equations with Lotus 1-2-3." The author demonstrates how to formulate and solve a famous macroeconomic model, Klein's Model I, using standard Lotus commands. The Gauss-Seidel iterative method is used to numerically solve the system, with a one-line Lotus macro written to test for convergence.

Another example of Supercalc use is to make projections of coarse grain production in foreign countries using population projections, real GNP growth rate, elasticities of consumption with respect to income, and growth rates of production. The spreadsheet format allows the analyst to change one item, such as an elasticity and have everything else recalculated. In this way it becomes easy to cross-check to see if implications of certain assumptions are reasonable.

A planned enhancement to this analysis technique is to begin to use the regression capabilities of a microcomputer statistical package, ABSTAT specifically. Regression of grain conversions over time can yield estimated elasticities, which can then be put back into the spreadsheet.

ABSTAT was acquired as a user-friendly package to do basic descriptive statistics and simple linear regressions. We have also acquired SPSS/PC, the micro version of the popular mainframe package. Many of our economists are accustomed to using SPSS for analyzing survey data and large cross-sectional data files such as those provided by the Census Bureau. To provide databanking of larger files of which portions might be analyzed using SPSS/PC, we recently licensed SPSS/X to run on our in-house minicomputer. SPSS/PC's ability to handle "portable" system files which can be uploaded and downloaded easily aids in forming an interface between the large and small computers. We will first apply this in analyzing the results of an in-house information-needs survey; complete questionnaire results can be stored on the minicomputer, with data for particular groups of respondents or selected variables downloaded to the micro for detailed analysis without having to be redefined.

We have two packages in-house that can perform more complex econometric estimation techniques: RATS (Regression Analysis of Time Series) and SORITEC. A domestic sugar model has been set up in SORITEC. Various estimations were performed, including OLS and two-stage least squares and Cochrane-Orcutt autocorrelation correction for each equation. The model was too large at 15 equations for SORITEC to do maximum-likelihood estimation of it, but the new version, when it comes, should be able to handle it. The model was simulated in SORITEC with the various sets of coefficients and also with various changes made to the model, for example perturbing an exogenous variable by 10%. SORITEC has a command to compare actual and fitted values, computing summary statistics to measure goodness-of-fit.

Because the model is somewhat large, it is run in a "batch" mode, with Wordstar used to edit the SORITEC program. The model has also been put up on Lotus 1-2-3 to experiment with the parameters. Graphwriter is used to output plots of results.

There is a free version of SORITEC called SORITEC Sampler which has capabilities of the main package up through two-stage least squares. It cannot perform three-stage least squares maximum-likelihood estimation or

handle nonlinear models. It produces nice screen graphics of regression plots including residuals, which can be dumped to a line printer (but not at present to a plotting device). While not of publication quality, the plots are very useful for analytical work. For example, as part of a farm production model, an equation was estimated with prices paid by farmers for feed as a function of corn price and the price of soybean meal. The residuals showed some problems; an autocorrelation correction was tried and the regression re-estimated. The new plot showed substantial improvement in the residual analysis.

Another analyst uses RATS to estimate import demand for wheat, corn and soybeans in four Asian countries. The 10-equation model has been run through OLS, instrumental variables and Taylor-series approximations, and he is trying to get around memory constraints (supposedly temporary until the new release of the package) to do seemingly unrelated regressions. The ARIMA time-series analysis capabilities of RATS were used in this project in determining how to average prices on a yearly basis, looking at the cross-covariances between prices and imports to decide on a lag structure.

RATS is also being used to estimate a Canadian grains and rapeseed model. Again, a spreadsheet, in this case Lotus, is being used to update the data and provide graphical output, as well as to simulate the results.

We have at ERS a number of other software packages for microcomputers to perform more specialized functions. GAUSS is a matrix programming language that allows you to write out an analysis the way you would write it mathematically. You can easily write down the estimation commands for the coefficients of a simple linear model, or the code for a complex statistical algorithm as it appears in a journal article. GAUSS does not currently come with built-in statistical routines but is planned to in the future.

Another program, TK!Solver, solves simultaneous nonlinear systems, again allowing you to express the equations similarly to how you would mathematically. A package called MUMATH solves mathematical problems symbolically and can take derivatives, etc. Especially useful in microeconomic theory, one can change coefficients or other aspects of a model symbolically rather than numerically and see the logical implications in terms of cross-relationships that result.

We even have some researchers who use small programs written in Basic to perform a specific statistical function, such as regression or the calculation of standard deviations or coefficients of variation, rather than bother learning how to use a more complete statistical package.

Finally, I would like to mention one macroeconomic model to which ERS subscribes, FAIRMODEL which is a model of the U.S. economy developed by Professor Ray Fair of Yale University and programmed for the IBM PC and XT. The model consists of 30 stochastic equations and 98 identities and is re-estimated quarterly. It can be used for forecasting, policy analysis, scenario development and as a research tool. An analyst can run experiments: change exogenous assumptions, enter adjustment factors, or exogenize an equation or block of equations, and view the results. An interface to Lotus 1-2-3 can be obtained with FAIRMODEL to use for setting up an analysis and deriving tables and graphs from the model output.

These have been only a few of the very many applications of microcomputers that we have in-house. The use of microcomputers has revolutionized the way our analysts conduct their research. In the area of econometric modeling, many more alternatives can be considered and assumptions tested in a much shorter period of time, taking advantage of the interactive nature of the software on these machines. Researchers who in some cases had little computer experience previously have become proficient with the easy-to-use and flexible software available on microcomputers, particularly spreadsheets, and seem to prefer this to the use of cumbersome statistical packages. However, now that better statistical software is becoming available, interest in it is growing. The economists I spoke with seemed to want to choose their own components of an analysis system - spreadsheet, statistical program, graphics package, word processor - and are concerned with having good interfaces so they can quickly move data from one program to another. Some problems with memory constraints and speed have been experienced, but hardware is rapidly improving to alleviate this. There are worries about having errors creep into programs, especially with spreadsheets that may not be well documented and might be passed from one researcher to another. These and security of data issues will have to be addressed now by analysts who perhaps had that taken care of for them in a mainframe environment, but this seems to be a fair trade for the ability to interact directly with their models and better understand what the data are saying.

#### **Reference**

Johansson, Jan-Henrik, "Simultaneous Equations with Lotus 1-2-3," Byte, February 1985, p. 399.

#### **Acknowledgments**

Many thanks to the following researchers who shared the results of their work on microcomputers: Walter Ferguson, Vernon Roningen, Michael Lopez, David Weisblat, Suchada Langley, Gary Lucier, Carlos Arnade, Larry Deaton, Clark Edwards, Paul Prentice, and Merv Yetley.

#### **Software Vendors**

AbStat  
Ander-Bell  
P.O. Box 191  
Canon City, Co 81212  
(303) 275-1661

SORITEC  
Sorites Group, Inc.  
P.O. Box 340  
Springfield, Va 22151

Graphwriter  
Graphic Communications, Inc.  
200 Fifth Avenue  
Waltham, Ma 02254  
(617) 890-8778

SPSS/PC  
SPSS, Inc.  
444 Michigan Ave.  
Chicago, Il 60611  
(312) 329-2400

Lotus 1-2-3  
Lotus Development Corporation  
161 First Street  
Cambridge, Ma 02142

SuperCalc 3  
SORCIM/IUS Micro Software  
2195 Fortune Drive  
San Jose, Ca 95131  
(408) 942-1727

MuMATH  
Microsoft Corporation  
10700 Northrup Way  
Bellview, Wa 98004

TK!Solver  
Software Arts, Inc.  
27 Mica Lane  
Wellesley, Ma 02181

RATS  
VAR Econometrics  
134 Prospect Ave.  
Minneapolis, Mn 55419

FAIRMODEL  
Urban Systems Research & Engineering  
2067 Massachusetts Avenue  
Cambridge, Ma 02138  
(617) 661-1550

**SPREADSHEET AND DATA BASE APPLICATIONS USED BY THE CROP REPORTING  
BOARD IN REVIEWING SURVEY INDICATIONS AND PREPARING PUBLICATIONS  
Gary Nelson, U. S. Department of Agriculture**

Our Agency, the National Agricultural Statistics Service, is responsible for gathering crop and livestock statistics for the Department of Agriculture. We make forecasts of the crop size during the growing season, and final estimates at the end of the year. We have a network of 44 field offices serving all fifty states. These field offices regularly survey thousands of operators of farms, ranches and agriculture businesses to gather information about their operation. Statisticians in our field offices assemble the information and make recommendations on such items as acres planted or harvested, yield per acre or the amount of grain that is in storage. They then send indications and recommendations to our headquarters office in Washington, D. C. where the data are assembled and reviewed, and U. S. estimates are set and published.

Our state offices are connected to a large computer network, the Martin Marietta Data System. The indications, recommendations and comments are submitted over the network to our office in Washington, D.C. We have several IBM PC/XT'S in our section, which we utilize extensively for summarizing data and weighting the data to give state, regional and national totals, as well as designing questionnaires and various other spreadsheet applications and some graphic applications.

One microcomputer application that we have developed is called the Grain Stocks Program. This program produces a report that we release four times a year, that shows the amount of grain that is in storage, both on the farm and what is stored off the farm. The report was produced manually in the past and we wanted to put it on the micros. In designing this application, we wanted a system that would: be easy to use, be menu driven, be able to download the data from the data base on Martin Marietta to the microcomputers; assemble the data, provide a means for making changes, provide us with summaries and camera copy that we could use to print the report, provide the ability to transmit the changes back to the data base at

Martin Marietta, and provide the capability to compute and print a balance sheet. The program uses a combination of Condor and SuperCalc3.

Another application of the micros is in tabulating and charting data used in making forecasts on the size of the various crops each month throughout the growing season. These forecasts are released on a specific day each month. Since the forecast of the size of the crop can have a definite impact on the prices, it is extremely important that strict security be maintained in compiling these statistics until the report is released to the general public. To insure that the data are kept confidential, we operate under a "lockup" procedure. The members of the Board review the data, read charts, and recommend a yield for each State and the Region. The Board then jointly agrees on a yield for each State to give the U.S. totals. The biggest use of the micros in this application has been to assemble the data to a Regional level and at the same time provide printed worksheets to the Board members for setting the estimates. We usually have less than one hour to prepare the data for review by the Board. I can enter the indications on the PC, and within about five minutes print out the spreadsheet with all the indications on it. In the past it would take almost one hour with two or three people doing the calculations and checking the totals manually to complete these tasks. Furthermore, these time savings permit extra time for reviewing the data and ensuring they are correct.

In conclusion, we find ourselves putting almost all of our calculations on spreadsheets, and even people who have little experience on computers are able to effectively use the micros. In most cases there has been a considerable time savings, coupled with improved data quality.

**MANAGER'S PERSPECTIVE ON THE ACQUISITION  
AND USE OF MICROCOMPUTER-BASED GRAPHICS PACKAGES  
Richard W. Hays, Internal Revenue Service**

The capability to display statistical data graphically as opposed to tabularly has been greatly enhanced with the advent of graphics software packages which can be used on microcomputer equipment. This paper summarizes the experiences of one small statistically-oriented organization, the Projections and Forecasting Group, a component of the IRS Research Division, in using microcomputer-based graphics to upgrade the quality and impact of its products.

**Mission of the Projections and Forecasting Group (PFG)**

Until 1983/1984, the Group's projection activities were completely mainframe bound. All consequential projections were performed at a remote IRS computer facility in a dumb-terminal, time-sharing mode. There was no graphics capability in this system. Even tabular information was difficult to extract in a format which was ready for camera-copy reproduction.

The introduction of 16-bit 10 MB hard-disk micros into the Group in early 1984 radically altered work processes within six months:

- Lotus 1-2-3 spreadsheet software was used to format smaller projection projects.
- A downloading capability was created so that large-scale computations could be done on a mainframe with numbers dumped into preformatted tables.
- A variety of different tables were created which allowed more rapid scanning for errors or problems in projections.
- Data transmission arrangements were made with key users so that data previously supplied in hardcopy only could be provided electronically, thereby facilitating further analysis by the user without data re-entry.
- Experiments with Lotus 1-2-3 graphics suggested that much could be done to present analytical information and projection highlights in pictures rather than words or spreadsheets. Graphic representation of data would expand the managerial and executive audience.

### Graphics Experimentation

Early experiments in presentations were done using Lotus graphics and a dot matrix printer. The Group found Lotus graphics satisfactory but limited in the quality of presentation both in terms of sharpness for reproduction purposes (a printing problem) and sophistication (a software problem).

The problem of quality reproduction was solved by acquiring a six-pen Hewlett-Packard plotter. Tests and discussions with other organizations showed eight-pen plotters to be too slow, too complicated and too expensive. The second problem, sophistication and flexibility of presentation, necessitated a software survey. A number of different packages were reviewed against survey criteria:

- compatibility with Lotus 1-2-3 files,
- menu structure,
- equipment compatibility, and
- memory demands.

Chart Star, software marketed by Micro Pro, Inc., was chosen at the end of this survey. Chart Star has a wide range of charts and graphics to choose from and is in all ways superior to Lotus 1-2-3 graphics. For example, bar graphics can be three dimensional; it has exploding pie charts, and a number of other options, all prefaced with easy-to-use menus.

With both hardware and software in place, the Group began to routinely use graphical data representations in its reports and documents. We discovered that once it was demonstrated what microcomputer graphics could do, demand for such presentations increased exponentially. Consequently, the Group added to its repertoire a software package called Statmap which permits shaded/cross-hatched representations of data on maps at the zip code, county, state and U.S. level.

## Some Observations on Impact

There is no question that microcomputer graphics have greatly improved the quality of and increased the audience for Projections and Forecasting work productions. Managers and executives are more aware of key trends and have tended to ask for additional data and displays on them. There are organizational impacts, however.

Search time - Finding the hardware and software which suits the presentation requirements of the organization requires time--staff time. Hardware and software specifications need to be reviewed, demonstrations arranged and procurement initiated. Getting the right technology for your needs requires carving out enough time from everyday work to do an adequate job of review.

Implementation - Training employees to use graphics software is not usually a major issue. However, selecting the right graphics to demonstrate the point in question is a more substantive issue. Doing so requires consultation and testing and adds to production time. The longer the review chain, the more frequent will be requests to alter graphic presentations or present data in some other manner.

Integration - Good graphics create their own demand. We found that top managers expect textual material with high data content to be graphically illustrated. We have not found software which does a good job of integrating text and graphics for camera-copy development. This means graphics and text are separately produced, then cut and pasted into camera copy. The result is that making changes becomes more difficult than simply making textual adjustments on a word processor.

Color - All good graphics packages can give video displays of charts and graphics in color. With a plotter, camera copy can also be produced in color as can overhead projections. The rub develops in moving from camera copy to production. Few organizations have the color xerography necessary to make color reproductions, although this may be coming. For the interim, graphic presentations need to be developed with a black and white final product in mind.

Competition - Good analysts quickly realize that graphic data representations help sell their products. Consequently, there is competition for the use of both equipment and software. If the organization has micros with either built-in or external hard disks, the equipment side of the equation can be solved by loading software into the hard disk. The number of software packages needed will depend on the volume output of the Group. In our case, during peak production, one package for five analysts seems to meet the need. Supervisors and/or reviewers need also to guard against "over illustration," a problem which can occur once analysts have seen the power of graphical presentations.

## CURRENT APPLICATIONS OF UNIX-BASED MICROCOMPUTER SYSTEMS

Brian Carney, U.S. Department of Agriculture

The situation in the National Agricultural Statistics Service (NASS) is unusual in that several of our microcomputers are based on the UNIX operating system. Instead of having just one user on a machine, we have multiple users and multiple tasks per user.

First, a little about what the Research Division of NASS does. There are three branches in the division: Remote Sensing, Yield Research, and Sampling Frames and Survey Research. The Remote Sensing Branch uses a UNIX system with exotic graphics hardware for satellite image processing. The Yield Research Branch is using a UNIX system for editing programs that are submitted to a mainframe computer. The Sampling Frames and Survey Research Branch, the group I am in, works in three general areas: nonsampling error research, area-frame design and construction, survey design and analysis, and statistical consulting.

Much of the work of the latter branch involves large datasets from our agricultural surveys and requires the use of a statistical package on a mainframe. We use SAS primarily.

Efforts to reduce nonsampling errors in telephone interviews is what led to our use of UNIX. We became research partners with the Center for Computer Assisted Survey Methods at the University of California at Berkeley. That group had been working on a system for computer-assisted telephone interviewing (CATI). Using the CATI system, we replace paper questionnaires with a terminal, and the interviewer can enter respondents' replies directly into the computer. Error checking is performed while the respondent is still on the telephone. CATI was developed and runs under the UNIX operating system.

For a while back in 1982, when we were just starting our work with CATI, we tried connecting the terminals over telephone lines at 1200 baud; and at that slow data rate, it takes a while to paint the screen, delaying the interview. Shortly thereafter, some multiuser microcomputers became available that ran UNIX; and the CATI system was ported over to them easily. The cost of the system was not too bad, something under \$40,000; so we were able to procure an inhouse system for a work group of about ten people. That was our introduction to UNIX.

Once we had installed the machines, it was clear we could do quite a bit more with them besides run CATI. What we have done falls into the category of analysts' support. There is a video display terminal on each desk, with access to the mainframe systems both interactively and in batch mode. Programs for analyses are written on the UNIX system, using the native full-screen editor, and are transmitted to the mainframe for execution. This avoids the cost of being online to the mainframe. By dialing in to other systems or having them dial in to the UNIX systems, we can transfer information on the word processors and PC'S in and out of UNIX. The electronic mail system has been very useful to the managers; and some of the material usually covered in staff meetings is now mailed to the staff using UNIX, and they can read it at their convenience. The electronic mail system extends to the research UNIX systems in Washington and the field.

Of course, UNIX has tools to facilitate programming, technical writing, and publishing; and these are widely used on our UNIX systems. All this capability is right there under UNIX; many operating systems are not so rich.

The communications capability of these systems is such that they are accessible from remote dial-in terminals in the same way as mainframes and minicomputers.

I mentioned the specific use of these systems for CATI research. The Agency, because of the interest in making CATI operational, has procured twenty to twenty-two UNIX-based machines for our field offices. Besides running CATI these will be used for direct data entry, transcribing the responses to paper questionnaires we still generate in field interviews.

Software costs have generally been lower since one package is purchased per system. The individual price is high, but is generally cheaper than PC software on a per-user basis.

I have mentioned office automation. All the analysts have a CRT on their desk, and prepare reports and analyses through that terminal. We can edit and review before the manuscripts hit hard copy. This can be a real time saver.

There are several spreadsheets available. A number of imaginative simulations have been done on spreadsheets under UNIX by our group. Database software is available, but is used now primarily for administrative purposes.

The statistical analysis capability under UNIX is a limitation right now. Probably one of the most complete statistical languages available is the AT&T Bell Labs S. However, it is very large and does not run on every UNIX system because of certain hardware and memory requirements. P-Stat and Minitab both run under UNIX, but would have to be converted to run under specific systems. SAS might run someday under UNIX, but probably not for a while. The SORITEC system that Linda Atkinson mentioned is available, too, and is good for econometric analyses. A system called UNIX/STAT is available for basic statistics and psychometric analyses. Because of these limitations, we do not do much statistical analysis directly on the UNIX systems.

Now a little about UNIX itself. Among its disadvantages is its size: it requires as much as twelve megabytes of disk for the operating system and utilities, of which there are hundreds. The system can appear to be quite complicated, and it is usually necessary to have someone available to help out with solving system problems. The commands are a bit terse, most only two to four characters long, and that can be a problem for new users.

Among the advantages of UNIX are its flexibility and power. You have an operating system that operates on minicomputers, mainframes and microcomputers and has the same essential capabilities across them all. UNIX is powerful because you can do extremely complicated functions with a very small number of keystrokes. The multitasking means that a user can have several programs operating simultaneously. When I use viewgraphs, I

set each one up interactively, then have the system actually draw the viewgraph in the background. While it was drawing, I could go on to the next viewgraph.

There are hundreds of utilities native to UNIX that are useful for the full range of tasks from text processing to database to programming. The UNIX hierarchical file structure is important for managing large numbers of files.

Some of the new UNIX systems feature displays with multiple windows that can run different processes at the same time. For applications requiring detailed graphics or typesetting, several new systems use bitmapped screens. What you see is what you get, even to the different fonts, special characters, and drawings. The results, printed on a laser printer, are quite good. It is not unlike the Macintosh, but with a more substantial operating system.

Decision support on a microcomputer is the idea of having all the functions an analyst needs for assembling, analyzing, and presenting data, in both graphic and text form. The systems are flexible enough to manage both text and graphics in the same files.

UNIX is also developing sophisticated networking to allow shared access to file systems, but with separate processors available to each user.

We have found the UNIX systems to be extremely useful because of their power and the wide variety of utilities built into the operating system. But we have been limited by the small number of statistical packages available under UNIX, and still rely on a mainframe for most of the statistical analyses.

#### **EQUIPPED FOR THE FUTURE?**

**Paul Dobbins, U. S. Department of the Treasury**

The Office of Tax Analysis (OTA), which is part of the Treasury Department's Office of Tax Policy, is responsible for three major functions. First, providing revenue estimates for the Administration for the budget and its quarterly reviews, as mandated by law. Second, providing on demand revenue estimates of tax proposals. Third, providing economic analysis of current and proposed tax legislation, often on very short notice. Timeliness is of the essence for the work of OTA to be of any impact during the sensitive, if fast-paced negotiations that are characteristic of mark-up whenever a tax bill is pending.

OTA specializes in what is called micro-simulation, which in our case is simply defined as modeling the responses of tax or household units to tax-law changes on an individual basis and weighting up the sample results to get population estimates. Our data is primarily tax-return information, but we are making increased use of other sources.

Even though our data files are relatively small samples, they are still large data sets and have made us largely mainframe bound. (Micro data does

not imply microcomputing!) But we have begun to use microcomputers to increase overall office productivity and, having seen the light, we are attempting to push our frontiers even further out to the limits of the possible.

We would ideally have a triad of computers/computer systems supporting our work. At the top or first level would be the individual microcomputer workstations bringing the burgeoning new wave of software development into the hands of tax economists and lawyers. Currently only our economists and computer specialists have microcomputers: the Z80-based CP/M machine, the Superbrain (TM), a fine machine in its day but certainly no longer in the mainstream due to memory and system limitations.

Our economists have benefited greatly from having these microcomputers. Most of the staff were quickly converted into very proficient users of Wordstar (TM) and Supercalc (TM), and several have demonstrated considerable programming talent. The programming staff has found its burden lightened somewhat by a transfer of focus by staff wherever possible from the mainframe to the micros.

But as described above, we are mainframe bound by the nature of our micro-simulation bread and butter. The mainframe is the third and fundamental level of our triad. What we're hoping to implement is a second level: a mini- or supermicro-based network linking the first and third levels together into a smooth, efficiently running system.

Why this is needed can be illustrated by a generalized paradigm of how OTA often does its work. A particular tax proposal will need to be reviewed and analyzed in the course of an afternoon. First, the appropriate mainframe simulator will be run and the results brought to the staff, who may then input the numbers into an analytical framework they have developed (e.g., in Supercalc (TM)). These modified results then may be inserted into a document residing on a third device, a stand-alone word processor. Finally, we have results, but not without considerable time being wasted simply carrying paper from office to office and re-inputting numbers at each step along the way. This is, I submit, an old story for many an office. We have duplication, wasted effort. And yet the very presence of microcomputers has made the process faster and more reliable.

A triad of micro-mini-maincomputers seems to fit our demands perfectly. The new Micro-Vax (TM) may very well fit into the second slot. The workstations could easily be IBM PC's and look-alikes or DEC PC's, while any mainframe fits the foundation (we're running on a UNIVAC 1100/81 series and will soon have an 1100/92).

Our ideal system would make it possible to share software and software tools at the local and network levels, and allow us to easily move text and estimates across and among all of its levels. The intermediate level also goes some way towards narrowing the software gap between mainframes and micros by providing considerable computer power and much of the latest software design. We can only anticipate the hardware advances that will narrow the gap even further.

Finally, it cannot be overemphasized that the introduction of microcomputers has fundamentally altered the way OTA does business. But perhaps the

greatest lesson learned was how much better we can do than our current partial solution. In an environment that features considerable interaction among many different specialists, a network unifying all computer assets seems to be the only way to go.

**CONCERNS ABOUT DATA INTEGRITY, SECURITY, AND ACCESSIBILITY IN AN ENVIRONMENT  
WHERE MICROCOMPUTERS AND MAINFRAMES ARE INTERFACED  
Dick Shively, U. S. Department of Agriculture**

**A. Background (History)**

The NASS is known as a data collection agency and reporter of agricultural statistics. In this line, a substantial amount of effort is directed toward list maintenance and data collection.

As the agency moved into single- and multiple-frame probability samples for more and more indications, large-scale computer resources became extremely critical to allow evaluation of results in a manner timely enough to meet the reporting schedules, as well as to support much more exacting requirements for list maintenance.

A large proportion of the field office (SSO) efforts are directed toward maintenance of the lists associated with their individual state, as well as collecting and analyzing their data. The NASS estimating procedure normally consists of each SSO preparing the indications and estimates for their individual state; then the Crop Reporting Board reviews all of these estimates to arrive at regional and national estimates.

Since many diverse commodities are estimated, and sample sizes are fairly large to cover the desired geographic areas, data conversion is a major task. For this reason, all of the SSO's are mainly equipped as remote-job-entry sites with high-volume data-entry equipment.

While the NASS has been a proponent of "generalized" software for some time, operations of this software at multiple sites required that installation and maintenance activities were duplicated for each site.

The NASS history in mainframe computing has progressed from each of the field offices (SSO's) being responsible for obtaining their own computer resource locally to the current approach of providing a single large-scale commercial time-sharing-network vendor who can provide adequate resources to satisfy all of the agency's requirements.

Until the bulk of national interest surveys were processed on the time-sharing vendor's equipment, it was sometimes difficult to determine what procedures were used to obtain indications and even more difficult to review multiple state outputs from different reporting formats. Even shared software required modifications to be used on different brands of hardware, and there was little assurance that these modifications would always provide identical results.

The introduction of microcomputers into the NASS offers some possibility of encountering the same problems recognized when local computer capacity was utilized. However, with care in selection of commercial software to ensure standardization where necessary, these devices offer substantial opportunities for improvement in personal productivity.

## B. Data Accessibility

The typical processing method for survey data in the NASS for national surveys is for the Washington, D.C., staff to provide general guidelines for the type and amount of data validation to take place, as well as the summarization techniques. The SSO staff provides detailed validation specifications appropriate to their state, taking into consideration any specialized local conditions. SSO personnel will collect and validate the data, and summarize to the state level. The D.C. staff consolidates all of the state level information into regional and national values. The majority of the post survey analysis is also accomplished by the D.C. staff, using the data from each SSO.

For this type of approach, the data values need to be available to people at widely-scattered locations. Storage on the microcomputer devices does not satisfy this requirement, since only one user at one location can access the data.

The same accessibility requirement also holds for Crop-Reporting-Board-released values. The D.C. staff reviews the SSO recommendations in establishing regional and national values. Following the Crop Reporting Board review during lockup, the state, regional, and national estimates are made public on a known date and time. These estimates, normally released at 3:00 p.m., must be immediately available to each of the SSO's to prepare reports emphasizing those items of local interest. In addition, many people outside of the NASS are allowed direct access to these published values.

Microcomputer usage in the NASS appears to be best adapted to play a support function, rather than providing a source of computational power. This includes primarily office-automation functions, such as word processing and spreadsheet analysis. Because of the volume of data, the stringent time constraints imposed on processing the data, and also the geographic distribution of processing, the data needs to be stored in a common repository. This allows each individual SSO accessibility to their own data, while still making the same data available to the review staff in D.C.

A recent "Viewpoint" column in INFOWORLD made the point that data security and microcomputer-enhanced productivity are incompatible. The relationship is very strong, in that the more stringent the security measures used become, the less accessible the data is with corresponding losses in productivity. This column's analogy is that security on microcomputers is similar to inventorying pencils and paper.

One current weak point with data accessibility, both from the mainframe and the micro standpoints, is a lack of communications ability. A reliable file transfer ability, consistent from machine to machine, and allowing usage of the strong points of both mainframe and micro, will enhance our data processing capability.

### C. Data Integrity

Storage of data on microcomputers in the NASS environment, for more than a temporary working basis, has a tendency to lead to integrity problems. This happens whenever the same data is stored in more than one location, whether it is on a mainframe or microcomputer, or in a file cabinet. Anytime that a value has occasion to be changed, unless all occurrences are simultaneously changed, someone will likely accept the wrong value as correct.

Version control is another name for data integrity, and when using individual stand-alone microcomputers this is difficult to handle. Each machine has its own copy of software, so whenever a new version becomes available every micro user must be provided with a copy. This is compounded by those machines having only floppy-disk drives, where the same software and/or data may appear on many diskettes. To upgrade to the new version, all copies must be located and modified.

The problem of maintaining data integrity on microcomputers is the same one we have been battling for years - a single copy must be identified as that containing the "correct" values, and any accesses must be directed to this copy.

Local Area Networks (LAN) provide some help in those situations where data is needed in a small area, where all users can be contained within the LAN. In this way, each user of the LAN can have accessibility to the same values, which will alleviate the integrity problem.

### D. Data Security

The most secure system is undoubtedly manual, although it may not be very productive. Microcomputers are not the place to store data that is sensitive, unless special considerations are made such as a locked environment and limited access.

Floppy disks are extremely easy to duplicate and just as easy to carry away from an office without detection. However, this is a human security problem and not a microcomputer security problem. An authorized person can just as easily remove a printout from a mainframe computer or pages from a record book in a file cabinet.

A "Jim Seymour" column in PC WEEK suggests increasing security by locking up diskettes when not in use, checking them out for usage, and encrypting the data stored on the micro. To me, this seems contrary to the usage of a PC as a productivity tool.

The main security consideration that should be given to microcomputer usage with sensitive data is the RF or radio-frequency emission associated with them. Using fairly inexpensive devices, these emissions can be recorded from many feet away from the computer and reconstructed into the data that was being processed. This can be solved with Tempest machines or RF shields.

In closing, I would like to say that I think that microcomputers are an excellent productivity tool. We need to be aware of their strengths and limitations when designing projects for them to accomplish.

## QUESTIONS AND ANSWERS

The following discussion reflects questions and answers related to the "Applications of Microcomputers" Session, which involved the Chair, Speakers, and certain members of the audience.

Mr. Steele: We have heard several interesting applications discussed here this afternoon, and the purpose was to give you some sense of the breadth of applications that microcomputers are being used for. We have seen applications that were very simple uses of spreadsheets, spreadsheet templates, and data bases, ranging on to econometric projections, graphics, and then the sophisticated time-sharing multiuser UNIX-based systems. We wanted to give you a sense of the kinds of things that are being done so that we can talk more about where we are going. What more can we expect out of microcomputers and what kinds of developments would we like to see in terms of hardware and software to achieve greater productivity from the use of microcomputers?

In that light, I would like to start off and ask several questions.

**Q1: Linda, you seem to be a strong proponent of the IBM microcomputers. Could you explain why?**

A1 (Ms. Atkinson): Being with the Federal Government, I am probably supposed to start with some sort of a disclaimer about products; therefore, my comments do not constitute a product endorsement.

When we first started acquiring PC's, it was done by our Economic Research Staff; and we had a proliferation of all kinds of machines including some that are no longer being made. It became clear that not only was this hard to support, but we were going to have problems in being able to move data from one machine to another and possible communications problems later should we want to network them.

Also, the situation that we are seeing now is that the newer software such as SPSS PC and SAS, are being developed first for the IBM machines and later, if at all, for other machines. So if you are on these other types of computers, you are going to lose some time during that waiting process. I would say by now that probably about eighty percent of our PC's are IBM compatibles or IBM.

**Q2: Brian, I see you seem to be a very strong proponent of the UNIX-based machines; and you give very strong arguments for, well, a good discussion of the relative strengths of them. We have heard a lot of sales hype about UNIX being the operating system of the future. Is that really going to happen?**

A2 (Mr. Carney): AT&T would certainly have us believe UNIX will be the operating system of the future. UNIX offers some special capabilities that you don't get on other ones. One in particular is software compatibility across UNIX machines. As an example, CATI software will run on practically anybody's UNIX box, anywhere. So, we have acquired a degree of vendor independence right there.

But, as far as hardware is concerned, say for example you want to network machines together; at that point you are down to a level where UNIX doesn't really do you a whole lot of good, because the individual vendors choose to implement UNIX differently on different types of hardware. There is some effort to relax those restrictions. Sun Microsystems has a hardware and software implementation of networking that can be done that is largely vendor independent. But that's not something in general that you can get and walk out with today.

UNIX suffers from being extremely complicated. As I mentioned earlier, you have to have an expert around or available when the machine goes down. Because of the size and complexity of the system, it does take the user quite a while to be productive.

In terms of the so-called popular software (Lotus, Symphony, etc.), none of that is available under UNIX. Whether it becomes available depends a lot on what AT&T can pull off with their market. They say it will happen, but we haven't seen it yet.

**Q3: Paul, as you look towards implementing your design of the future, having this triad or three levels, do you anticipate making them all one brand or one vendor and one standard software line; or what kind of problems do you envision in the interchange of data between packages?**

**A3 (Mr. Dobbins):** First of all I would say that we don't have enough experience at the moment to say exactly what it is that we will have a year or so down the road. I personally would look forward to having more flexibility and not essentially going toward one standardized system or one standardized software. We ultimately might want to cut bait with our mainframe. As the minis become more powerful and we have super-micros on the desk, we may begin to use the mainframe for initial processing and then download a data base to one of our PC machines.

I am looking forward to experimenting for a while before I will be able to say too much more.

**Q4: Rick, you mentioned that you had some problems with output devices. Could you expand on that a little? What problems have you encountered, and what could the vendors do to solve some of these?**

**A4 (Mr. Hayes):** What we are looking for is a good vehicle for integrating our text with our pictures or graphics. We haven't found that and we still are looking. One of the things we find is that there are a tremendous number of products out on the market. We are a small shop and we are doing our own research. So far we haven't had any luck with Text/Graphics integration. Once you start experimenting you find that you can spend a lot of time looking at various packages.

The other thing that you will find is that people will suggest different software you can look at, you can try. Each of these experiments takes time out of your production so there is a tradeoff between finding something that will work for you and finding the best and latest product. Once you have found something which works for you, I think you should stick with it for a while and find out its complete capabilities. When we talk about output devices, we are having problems with finding print devices that will handle

graphics, pie charts, diagrams, as well as text, at a reasonable speed. I think most of the products presently available have problems with integration of text and graphics. It is not an insurmountable problem, but it slows you down.

Q5: Dick, you expressed several concerns about data security when interfaced with mainframes. Does that mean that you think people shouldn't have microcomputers hooked to the mainframes or that we should lock everyone's door or what?

A5 (Mr. Shively): All I was saying on that was that the microcomputer has a problem with security. Mainframe computers are very well adapted to securing data, securing devices, making things pretty secure. Microcomputers, by design, are a productivity tool. If we try to add security to it beyond what is necessary, we have reduced a lot of the productivity gains that we have possible there.

Mr. Steele: At this time I would like to open the floor up for general questions. Please stand and identify yourselves by your name and agency before asking the question.

Q6: This is a question that is appropriate to organizations that haven't yet capitalized on current technology and are looking to get into the business. What is the appropriate level of computer power for economic feasibility forecasts and statistical work? I wonder if anybody can comment on the relative merits of waiting until 16-bit-microprocessor technology is developed.

A6 (Mr. Hayes): As it turns out, we recently had a contractor come through and evaluate our operation in terms of what ought to be our computer configuration in the future. Basically, their conception is that anybody who is working with any sizable base is not going to get along without it. Their suggestion is that you use the mainframe for your heavy duty computations and as a storage device to deal with some of the security problems we talked about here. They then are suggesting to us, at least, that we look now to microcomputers, 16-bit machines that can network into the mainframe and be used either in the stand-alone capacity for test purposes to try out graphics, to upload and download, or as remote terminals.

I know we are going to maintain our mainframe capabilities while doing as much as we can on micros, because micros are much simpler and flexible to use by analysts. Where we need to spend time is on gateway architecture that links our micros to the mainframe and allows analysts who are not experts in programming to get in and out of the mainframe and get data in and out of software packages.

A6 (Mr. Carney): In the UNIX-based workstations that I am familiar with, the real 32-bit technology comes on a single chip with a powerful box. I think people are looking towards Motorola 68020 which should be in production sometime in the Fall, 1985. It takes a little while, not too long, under UNIX for the software to catch up with it. It may be a while before it's really fully mature as you are looking for right now.

You've heard this old song before, but you really want to find what software you want first, and then figure out what sort of boxes you can afford, that it will run on. You really are looking for productivity after all.

Q7: Each of the panelists has described a decentralized system, especially one using spreadsheets and/or spreadsheets with something else. How do you insure the integrity of the data that you are using; and second, how do you insure that whatever statistical standards may exist in your agency are being followed in those decentralized systems?

A7 (Mr. Carney): I can talk about it a little bit in the research environment. Basically, we always have to use the data on the mainframe as the benchmark data. That is the correct data, and anything we pull off has to be pulled off checking protocols; and we can't change those numbers.

As far as the statistical standards are for the research unit, you pretty much have to depend on the review process, review by our peers.

A7 (Mr. Shively): I second Brian's statement. We basically consider the mainframe data to be the official source unless some special circumstances exist where there is no need for it to ever be on the mainframe. But for any data that is shared or that is nationwide in scope, the mainframe is the official source of control. If you pull a copy of that to your micro to run it through a model, you are working with data that at that point in time is not official copy. It may be a copy of the official and you can use it for your model or plan -- anything like that -- but if you want to go to publication with it you need to go back to the mainframe for the official copy.

Q8: This is a comment. We are using more and more graphics lately. There are some packages on the market which were released recently that will capture the picture of the U.S. map -- then you can enhance that map. You can add titles, text, whatever you want to these graphics. The program works in the background -- one is the graphics partner. You can call it like the Sidekick package. It captures the picture you have on the screen and you can enhance and modify the picture. You have the integration of text and graphics. I suggest that you try SMART; they have the software package. Also, take a look at GEM which is just coming out by DIGITAL Research.

Q9: I have a question for Linda Atkinson. I think you mentioned a spreadsheet that ran for forty-eight hours? Did you have an 8087 (mathematical functions) processor.

A9 (Ms. Atkinson): Yes we did. It is a very large model and this is the simplified version of it on the micro, and yet it took that long. It's very large and it ran very long; but yes, it was an 8087 chip in there. They are hoping that the AT is going to improve the situation. If not, they may not be able to ultimately move to the micro but will need to use the mainframe as well for that model.

Mr. Steele: I have one anecdote that I would like to share with you about computers and applications to computers. I became involved with microcomputers in 1978. By 1980 I had most of my functions already automated in my office on the computer, and I called up one of my colleagues who was Secretary of the Crop Reporting Board to get some information from him on when one of our employees had last been in on the Crop Reporting Board. He said, hold on just a minute, I need to check my data base, and within thirty seconds he had an answer back for me. He read back all the

times that this guy had been in and when he was next scheduled and I was really impressed. I couldn't believe how quickly he had all of that information. I didn't have a nice data base like that, so I asked him what kind of data base he was using, and he said, "file cards."

Certainly, the purpose of telling that anecdote is to illustrate that there are certain applications that are best left to a manual procedure, and then oftentimes I encounter people trying to automate procedures that aren't well defined manually. I think that any time people try to automate procedures that aren't well defined manually, they are expecting magic; and they usually end up with a lot less than what they are asking for.

Ms. Atkinson: I would like to make one comment about acquiring statistical software. If any of you are looking for software or a good source of reviews or products that people have tried, there are at least two electronic bulletin boards that I am aware of. Capital PC users group is a special interest group for statistics. Charlie Hallihan who is one of the chairs of that group has left some information on the desk outside which will tell you how to access their bulletin board or attend their meetings.

There is a SAS users' group, even though SAS is not yet on the micro. This is a group right now who likes SAS and also uses micros. I guess they like to discuss their applications of getting data back and forth from SAS. They have a bulletin board also.

A paper was presented on that at the SAS users' group meeting last month which I think had the phone number of the bulletin board. Otherwise, you can contact me. As I said, there is software available on these bulletin boards that people are willing to share. There is a SAS macro on the SAS users' group bulletin board.



## SESSION ON EXPERT SYSTEMS

### SESSION SUMMARY\*

Both the DATAPLOT and Editing and Imputation systems described here were not developed by computer scientists or knowledge engineers but by subject-matter specialists who were presented with new tools to assist them in improving their jobs. Although "expert system" tools and techniques were developed by a community of researchers who happen to call their field "Artificial Intelligence," the tools and techniques can be considered to be useful in their own right without the necessity to call the result "Artificial Intelligence" systems. In fact, there is good reason to say that none of the existing expert systems is truly intelligent or even expert. A true expert has the ability to learn new rules in his specialty and to apply common sense reasoning in cases where specific rules don't happen to reside in his "knowledge base." Both "learning" and "common sense reasoning" are areas of artificial intelligence research in which there are only a handful of active workers and in which progress has been slow. Contemporary "expert systems" neither learn nor exhibit common sense behavior when it is warranted. But, as a set of tools and techniques, expert system technology has proved to be useful for some specific applications. We have seen two examples of such applications today in Mr. Filliben's DATAPLOT system and in Mr. Greenberg's edit and imputation software.

I think it is worth noting that these successful expert system examples were done by mathematicians and statisticians, rather than artificial intelligence specialists or even computer scientists. Less than two years ago, there were dire warnings that expert system techniques could not be generally applied because there were so few PhDs being granted to people who had specialized in artificial intelligence research. What we are finding though, is that the techniques important for developing expert systems can be taught to people in other specialties. In fact, many organizations (including the Digital Equipment Corporation and IRS), given the choice of training artificial intelligence researchers in application domains or training subject-matter specialists in the tools and techniques of artificial intelligence have opted for the latter. The Bureau of the Census and the National Bureau of Standards show that good subject-matter specialists are perfectly capable of learning the techniques without any deliberate training program by their agencies.

One of the reasons to have a panel on expert systems at a conference on statistical uses of microcomputers is that such systems as DATAPLOT can be adapted to personal computers as soon as the personal computers are powerful enough to accommodate them. To expand on that theme please note that

---

\*Norman Glick, National Security Agency

today's personal computers are already much more powerful than the "supercomputers" of the early 1960's. We are guaranteed, given what can already be seen in computer engineering laboratories, that the inexpensive personal workstations of the future will be powerful enough to accommodate the kinds of systems that need mainframe computers today. But the existence of that future power for the benefit of an individual will make even more important some of the research that hasn't been discussed today but is part of what the artificial intelligence research community is concerned with. The ability to provide a user model to accompany an expert system addresses some of the points made in the talks and the question period today. Users do have different levels of sophistication and expertise of their own. We would like the system to accommodate to the needs of the user, even to adapt to the changing expertise of a single user. The same person who might need substantial help in using a system for the first few times, might ultimately consider verbose assistance to be a nuisance. The ability of a system to adapt to the evolving needs of such a user is a subject of active research in the artificial intelligence community today.

Since it was announced that this session might be on "pure fantasy," and since what we have heard from the Bureau of the Census and the National Bureau of Standards has been on eminently practical systems (whether they are called expert systems or not), perhaps we should end with some speculations that some might consider fantasies. One class of artificial intelligence research that promises to have relevance to statistical systems of the future is automatic programming. Both the editing and imputation and the DATAPLOT systems required that the statisticians and mathematicians write programs. Whether they were intentionally building expert systems, unconsciously building expert systems or simply writing a program to assist in statistical analysis, they needed to provide significant detail about how the computer should do what they wanted. If sufficient expertise of the programming art can be captured in an expert system and can be combined with sufficient expertise in a particular domain, even one of the domains we've heard about today, then the combined system might permit a user to state what he wants done, rather than the details of how he wants to do it, and a program to perform the job could be generated automatically. To a modest extent, so-called fourth-generation languages provide existence proofs of such systems today. These fourth-generation systems work in very limited domains (e.g., payroll and inventory control), but there is substantial research aimed at increasing the set of applications for which such approaches are practical. Some even see this class of activity as the future of software engineering. Please note that some differences exist in the life cycle of standard software relative to the life cycle of "artificial intelligence" software. It is clear that current software engineering techniques will not provide the quantity and quality of software required in the future. More statisticians, mathematicians, and psychologists will need to tell computers what they want done in the future without computer-specialist intermediaries. Let's hope the automatic programming "fantasy" becomes less fanciful so that, in the future, more subject-matter specialists can be their own "knowledge engineers" rather than to be dependent on programming specialists. Statisticians shouldn't have to spend inordinate time learning the details of how to use specific computers when their talent is to apply their mathematical and statistical knowledge.

**INTRODUCTION**  
**Terry Ireland, National Security Agency**

It's possible that the organizers of this workshop, and I was one of them, wanted to have one session on pure fantasy and this might be it. Building expert systems that model in software the behavior of human experts, and evolve in a natural way so you can more clearly understand the expert, is so clearly impossible that there must be--and is--an unlimited amount of high-priced advice on how to do it.

Some statisticians may argue that random sampling and surveying procedures on computers can already model the experts, so we really, perhaps, have two questions:

What do we mean by "an expert"?

If we know what an expert is and if we have one on hand, how do we go about modeling his behavior?

In order to give some more practicality and reasonableness to this presentation, we have made absolutely certain that none of the speakers is a computer scientist. However, they are skilled developers and users of software systems, and they have built expert systems.

George Lawton is a psychologist with the Army Research Institute. He has an interest in systems that support the interface between human factors and computer science. He will give the introduction to expert systems.

Jim Filliben is a statistician with the National Bureau of Standards who has an interest in systems that model and support statistical expertise. In fact, one of his software systems is said to be the most requested piece of software from NTIS.

Brian Greenberg is a mathematician with the Bureau of the Census. He has an interest in expert systems for data editing and imputation.

Roughly a year ago I gave a talk on expert systems -- an abstract talk because my practical knowledge was limited. After the talk, Brian came up and observed that he wasn't sure about the jargon that was being used, but he felt that he had built an expert system.

The Rapporteur is Norman Glick. He and I are both computer scientists and he may try to have the last word. Mark Winer, an economist from the Office of Management and Budget, is our Discussant who will keep us honest.

Computer scientists are trying to create tools for the development of expert systems and to make them commercially available. They are also trying to give the impression that they are the most skilled at eliciting information from experts. Thus, the once humble programmer now calls himself a knowledge engineer. Ultimately, a knowledge engineer is a person (sometimes statistician, psychologist, or mathematician) who makes the most substantial effort to understand and model the expert.

**EXPERT SYSTEM TUTORIAL**  
**George Lawton, Army Research Institute**

A number of things have happened recently that would lead me to change some of the things I say today had I the opportunity to do so. Fortunately, some of those things are available to you through your local newsstand.

One was the publication last month of a magazine called PC, not to be confused with PC World and PC Junior, which has a section describing a number of proprietary software packages which are available on microcomputers for developing your own expert systems. Anyone who is inclined to go out and build an expert system may look at this article and review the software.

The other was my attendance last week at a conference at Bell Labs which brought together a number of computer scientists and statisticians who were all interested in what I am going to be discussing here this afternoon. I was surprised to find that there are at least sixty people from the United Kingdom and the United States and at least a couple of European countries who are interested in this subject. No less than John Tukey of Princeton University believes that this is the next wave of software for statistical applications. It seems to me that this is something of a coming concern.

Expert systems come out of laboratories for research in artificial intelligence. Artificial intelligence, as I think everybody in the world now knows if he reads the popular press, is a line of research developed at MIT, Carnegie-Mellon University, and Stanford University in particular, concerned with building computing machines which will emulate high-level human cognitive capabilities.

In the earliest incarnation of artificial intelligence, primarily at Carnegie-Mellon University, researchers tried to develop very powerful general-purpose problem-solving algorithms which would give a user appropriate support in tackling a problem that a human expert could solve. Those programs were largely failures. As a consequence of those failures, research in artificial intelligence has converged on one organizing theme in the past decade: to be intelligent, computer programs have to be able to access large bodies of knowledge. It isn't their deep problem-solving capabilities that make people intelligent; it's the fact that they know a lot about the world in which they operate, and so it must be for programs.

An outgrowth of that discovery is a type of computer program which is essentially nothing but a large collection of knowledge and a relatively simple mechanism for accessing that knowledge and using it to solve various problems. These programs are called Expert Systems.

I intend to talk about what expert systems are, about how their software is written, about the software techniques that expert-system developers use, and finally about what statisticians might want to do with expert systems.

## ARCHITECTURE BASED ON THREE SEPARATE MODULES

1. KNOWLEDGE BASE
2. INFERENCE MECHANISM
3. USER INTERFACE

*Display 22.*

Following good programming practice, expert systems are modular and they usually have at least three fundamental modules. The first is a collection of facts and rules in a knowledge base; the second is a relatively simple program evaluator we call an inference mechanism. The third is the interface with the user which gives the user the illusion that he is dealing with something that's intelligent (see Display 22). That's what makes the machine capable of passing what we call the *Turing Test*. This test is very simple. If a person acting as judge asking questions cannot tell the difference between a machine and a human any more frequently than he can tell the difference between a man and a woman, then the machine must be considered intelligent. Of course certain rules prevent the obvious shortcuts (for example, a terminal should be used to ask and to answer the questions). And, of course, the respondents (woman, man, machine) are not required to tell the truth.

In the knowledge base we have some knowledge, and it has to be represented in some form that can be used by the computer. Almost every expert system that I have had occasion to study uses one basic knowledge representation (possibly supplemented by some others): some kind of conditional structure we call production rules. The fundamental idea of a production rule is strikingly simple. It has two parts. The first asks if something -- some state of the world -- is true. The second part takes some specified action if that something is true. Programmers know them as if-then constructs in programming languages. In conventional programs they are usually scattered throughout the program text. In a knowledge base they are collected together into a list of rules.

### 1. Production Rules

There are two classes of production rules.

- A. Situation-Action Rules (which are essentially data-driven procedure invocations)

e.g., If the data are skewed, then call a re-expression procedure.

- B. Conditional Assertions

e.g., If the case has  $V1 = 10$  and  $V2 = 20$ , then it is an outlier.

*Display 23.*

Production rules really break down into two different classes. One of them is something we call a Situation-Action Rule. It's essentially a pattern-invoked program. The other is really a conditional assertion (see Display 23).

Most expert systems are based on one of those two kinds of production-rule systems. There are at least two other ways of structuring knowledge that are widely used in expert systems.

*Conceptual Networks.* First, you may have a large number of concepts to represent in the knowledge base, and the concepts might be related to each other (for example, a class-instance relationship or a set-subset relationship). There are species of animals, and each animal species has members. It doesn't make any sense to represent all of the features of each of those members, so they may be organized into conceptual, often hierarchical, networks which show these relationships.

*Frames and other structured objects.* This is really an extension of the idea of a record with the record containing not only the usual type of data, e.g., name or age, but also other records as data. Moreover, the data can be specified as a computation, e.g., if you know the person's year of birth and you know the year, you can compute age without storing its actual value. Frames also contain default information to be used or computed when the required data is missing.

#### 1. Forward Chaining

e.g., If P then Q;  
If Q then R;  
If R then T;  
P therefore T.

##### explanation

By P infer Q; by Q infer R; by R infer T.

#### 2. Backward Chaining

e.g., T if R;  
R if Q;  
Q if P;  
P therefore T.

##### explanation

To show T, first show R; to show R, first show Q;  
to show Q, first show P; P is true.

*Display 24.*

*Inference mechanisms.* Because we are talking primarily about knowledge that is represented as conditional structures, we must have some logical reasoning process to make use of them. For example, we can use some of the

basic rules of logic: if we know that Proposition P is true and if we know that Proposition P being true implies Proposition Q is true, then we can conclude from these two true statements (one a fact, the other a general reasoning procedure) that Proposition Q is true. Or, we can reverse the process, starting with a goal to show Proposition Q is true and reasoning backwards, looking for a sequence of statements that could bring us to the desired conclusion. Again, most expert systems use some variation of the first and second form of reasoning shown on Display 24.

*Propagation of uncertainties, statistical reasoning.* A third form of reasoning, which we will come back to in a minute, attaches to each conditional structure something called a certainty factor. Statisticians might think that the certainty factor might be a probability because some vary between zero and one or maybe a correlation because others vary between minus one and plus one. In general, they are not that well motivated -- they are ad hoc. They are just numbers that somebody pulled out of a hat, saying this is how certain I am that this fact is true. The question is, how do they get propagated through a sequence of inferences? This is a difficult problem about which there has been much recent discussion.

*Inheritance of Properties:* Again, we can represent objects in terms of a network of class relationships. By default, certain things may inherit properties from their class. If Fido is a dog, then Fido is warm-blooded, because Fido is a dog and dogs are warm-blooded.

*Heuristic Rules:* In certain cases, no straightforward and logical procedure may apply, in which case you may apply what we call a Risk It Rule. It's a rule of thumb that says "if we don't know any better, do this."

*Meta-reasoning:* Last, but not least, an expert system may have rules about rules, a form of reasoning about process often called meta-reasoning. It deals with reasoning about the representation of the problem. All of these methods have been incorporated into the handling of expert systems.

#### USER INTERFACE

1. Read user input.
2. Provide user with useful output.
3. Explanation facility, to give the user a useful trace of the program's inferences.
4. Knowledge-base input and editing.

#### Display 25.

This is the most important part of an expert system. The distinguishing feature of an interface for an expert system, and this is a well-designed interface, is Item 3 (see Display 25). Expert systems, unlike other computer programs, explain the conclusions they reach. I would say there is no other really necessary feature for an expert system than this ability to

explain. Previously, I showed several logical forms for the reasoning process. They provide examples of what an explanation might look like.

There you have a crude expert system. They are relatively unfriendly. They just say "by this rule, I infer this; and by that rule, I infer that" and so on, until you get to a conclusion.

Other systems are much better at knowledge representations, including diagrammatic representations of good inference and explanations in good English (see Display 26).

#### **Traditional Software Engineering -- Linear Program Development**

1. System requirements
2. Software requirements
3. Preliminary design
4. Detailed design
5. Code
6. Debug
7. Test
8. Use
9. Maintain

*Display 26.*

How do you go about building an expert system? The methodology that most people use is a little different from the methodology you might have learned in a basic programming or a software engineering class. Display 27 shows the steps that a conventional programming text might tell you to follow when building software. By the way, expert systems are just enlarged software systems.

#### **Alternative Approach**

##### **Cycles of Progressive Refinement**

**Preliminary Requirements**

**Preliminary Knowledge Engineering**

**Prototype I**

1. Design
2. Coding
3. Debugging
4. Testing

**Prototype II**

1. etc., etc., etc.

*Display 27.*

This is an alternative list that is used by most people who have developed expert systems. Rather than starting from the specifications and going step by step through the program development, requirements analysis and the rest to a final software product, expert-systems developers follow an iterative process which begins with a small program that is written and tested, then elaborated, and written and tested again. In fact, the programming language used in this development methodology was designed to support the iterative and experimental development of software. The ability to express your ideas in a high-level flexible language enables the programmer to develop rapid prototypes or models of the system he wishes to define.

LISP is the language of American artificial intelligence research. Notice that I said research. It's the language of artificial intelligence research. That may not mean that it's the best language for artificial intelligence implementation. It is a functional programming language. That means that programs written in LISP are functions that can be passed around just as ideas are passed around and used where appropriate. They are more like mathematical functions in the sense that they have the mathematical properties that you associate with function, rather than properties that you would associate with FORTRAN function. That's a formal statement that I don't really want to defend any further than to say it really is true.

#### **Four Basic Components At The LISP Top Level**

**(print (eval (read)))**

The first three are in the top-level loop of the LISP interpreter.

1. **Reader**
2. **Evaluator**
3. **Print**
4. **A table of LISP objects which serve as a data base.**

*Display 28.*

LISP provides a series of operations that you would have to make for yourselves if you were going to write a program in something like FORTRAN. When you invoke LISP on a computer, you are invoking an endless repetitive loop which looks like this. That is a top-level interactive computing environment from which you can either ask for a computation or define a new function, very much as you would interact with another person. Sometimes you are computing values or ascertaining facts; other times you are developing new ideas. Both activities are done in the same environment. The innermost expression is ready and waiting for you to type something into the terminal which it will read. Then there is the high-level functional evaluator which knows how to evaluate any well-formed expression in LISP. Then it will print out the results of that. It continues to go through this loop. It is like a conversation (see Display 28).

Invoked by the functions in this loop are three programs: the LISP reader which includes both the ability to scan the characters entered and to format them into a LISP expression, the LISP functional evaluator which can evaluate or compute the expression entered, and the LISP printer which knows how to format and print the results of the evaluation. Moreover, the LISP reader also stores information about the names or identifiers read in. Most of this information is stored in a table that holds rules, values and names. This table is useful as the data base. It is more than a simple table produced, for example, by FORTRAN.

Why would you want to use LISP? Because LISP is interactive, you can write a program and see how it works almost immediately. Compilation is unnecessary. Its modularity enables you to write small segments of code in the form of functions, checking each one as it is written. LISP doesn't require declarations, although good programming practice suggests they be included in the final product. This enables you to develop functions quickly for experimental use. LISP dynamically allocates whatever kind of data structure you want to use. That means when you call in a function, LISP will make it immediately available to you. This means LISP handles all storage allocation for you, allocating it when needed, cleaning it up when you are finished with it.

Because there are no differences in LISP between programs and data structures, LISP can be represented as lists just as if it were another data structure. Therefore, it is easy for LISP to reason about or deal with its own functions just as humans examine their own procedures. As a consequence, LISP provides sophisticated tracing and debugging capabilities.

## **PROLOG**

**Prolog is based on a general-purpose pattern-matching and inference or theorem-proving mechanism called unification resolution. It is based on a formalism from symbolic logic called first-order predicate calculus.**

### **Basic Components:**

- 1. Read and Print**
- 2. Procedure evaluation based on unification-resolution, implemented as backtracking search.**
- 3. A data base which contains the definitions of procedures and the facts needed by the program.**

*Display 29.*

PROLOG is a sort of second-class language for artificial intelligence research in the United States, but it is gaining adherents (see Display 29). It has a big following in England and in Europe. PROLOG stands for PROgramming in LOGic. The idea is that PROLOG is a language based on a subset of first-order predicate calculus called Horn clauses. It makes use of inference procedures used in proving the correctness of logical

statements. It allows you to write computer programs simply by making statements about what you want to be true in a clause form.

PROLOG provides many programming structures you would otherwise have to build. It can read program goals or procedures and print out the results for you. It's interpretive and it gives you all of the facilities of allocation of storage and reformation of unused storage in a manner similar to the support environment in LISP.

You do not need LISP or PROLOG to build an expert system--but it sure helps. The two speakers who are immediately following me are going to talk about expert systems they built in FORTRAN.

What's important is to identify and to make use of the distinguishing features of expert systems: the abstractions and program structure.

How do we use expert systems in statistics? This is a review of suggestions discussed at the conference on Expert Systems and Statistics that I attended last week.

First of all, I don't know how many of you are statisticians; but if you are, your knowledge, as distinct from the knowledge of laypeople who come to you for consultation or as distinct from pure subject-matter knowledge (for example, economics), consists of a collection of strategies for working with a set of data. And those strategies could probably be readily represented in the form of an expert system which knows what test to do next. Parts of the data need to be cleaned up. One of the most active areas of research that I have found is the specification of individual statistical strategies in the form of expert systems, either as interfaces to existing statistical acronyms like S or SPSS or SAS, or in the form of a complete statistical system. Nobody wants to suggest that you can improve upon the capabilities of these statistical packages. What you may want to suggest is that it may be possible to improve their usability by adding something between the user and the package.

Another area of research concerns reasoning with uncertainty. Statisticians have something to say about that. I mentioned the adding of certainty factors to knowledge bases in most expert systems. An active area of research is to determine how those certainty factors should be propagated through a rule system and how certain conclusions can be based on uncertain knowledge.

Existing statistical software deals only with statistical ideas at the lowest level. It provides code to do things like least-squares fitting and so on. We want to use the ideas of expert systems to move software up one more level to deal with the abstract ideas of statistical strategy--the choosing of statistical methods and selective analysis of data in light of these methods. Our success in this area depends on the development and use of modern programming languages and on the development and use of expert system models.

## AN EXTENSION OF STATISTICAL SOFTWARE TO EXPERT SYSTEMS

James J. Filliben, National Bureau of Standards

The outline for this talk falls into four general areas. We are going to be talking about the real relation of an expert system to a particular piece of software, namely, DATAPLOT. I am going to speak first about DATAPLOT to show how the expert system can be described with respect to DATAPLOT. We are going to be talking about the general structure for the intelligent subsystem -- expert subsystem -- and DATAPLOT and go into the interpretation of conclusions mode and the analysis guideline mode. The last mode deals with providing a guide for carrying out data analysis. We will go through a particular data problem.

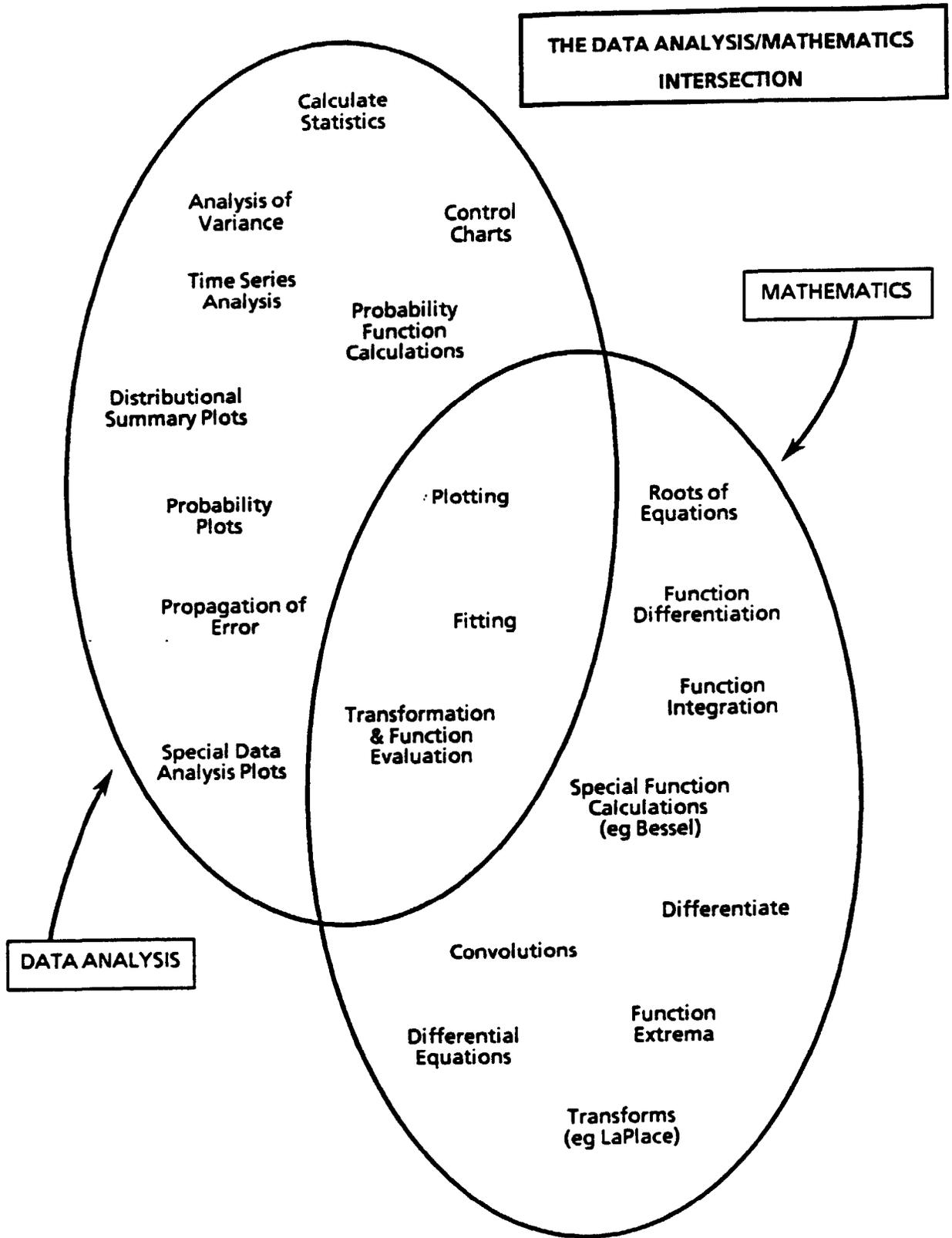
DATAPLOT is a high-level interactive statistical system with its own language, a high-level language with English-like commands. It was designed at the National Bureau of Standards (NBS) in 1977. The National Technical Information Service (NTIS) has been distributing it for the last three years. The software is written in FORTRAN. The cost is \$1200. It is the most heavily distributed software of its type at NTIS. It has been installed at about 200 sites. Next year it will be the most heavily distributed piece of software, period. Its primary capability is graphics.

That means it can run on Tectronix, HP and various other graphics terminal devices and on a variety of mainframes. It has both analysis graphics and presentation graphics. There are extensive additional capabilities in graphical data analysis and nongraphical data analysis, modeling and fitting, mathematics and diagram graphics.

At the National Bureau of Standards (NBS) we are interested in modeling and fitting data. In particular, we are often interested in fitting nonlinear models. Moreover, we make extensive use of applied mathematics and diagram graphics. By diagram graphics I mean the construction of schematic diagrams. The NBS is an engineering and scientific research organization. We have people that like to make schematics. We spend our time making schematic diagrams and charts. This component of DATAPLOT supports the automation of that work.

There is a heavy emphasis on data fitting in order to test underlying assumptions. The graphical displays are important because they provide insight into the underlying structure. Insight is important if you must go into court, for example, and defend your understanding of mechanisms at work in the data you have analyzed. Three notable cases that have arisen in our area are the analysis of the draft lottery, the argument over the use of daylight-saving time a couple of years ago, and data concerning the Alaska Pipeline. Graphical analysis was a critical component in those projects.

Display 30 shows the structure for DATAPLOT. It is a data analysis activity on one side and a mathematics activity on the other side. Three common activities common to both are plotting, fitting and various transformations and function evaluations.



Display 30.

Display 31 shows the typical commands you can issue to DATAPLOT. They support plotting (commands 1, 2 and 4), fitting (commands 3 and 5), and function evaluation (command 6).

#### TYPICAL COMMANDS

1. PLOT X Y
2. PLOT EXP(-X\*\*2) FOR X = -3 .1 3
3. FIT Y = A+B\*EXP(-ALPHA\*X)
4. BOX PLOT Y X
5. ANOVA Y X1 X2 X3
6. LET A = ROOTS SIN(X\*\*2)+EXP(-X) FOR X = 0 TO 5

Display 31.

Displays 32 and 33 give examples of the display capabilities of DATAPLOT.

All the graphics shown can be generated with any sort of system. Whether you have TECTRANIX, Spot 10, IGL or any graphics terminal, the important question is how long does it take to generate the graphical display. If it takes more than thirty seconds to a minute to do so, we lose the continuity so important to human-machine interaction. In data analysis the only concern is finding underlying structure, and getting insight. When generating graphics gets in the way of the objective of finding underlying structure in data, we lose control of the analysis. Thus, the utility of graphics software is measured not in what it can do per se, but rather in how easy it is to do it -- how easy it is to understand, write, modify and communicate the instructions.

The DATAPLOT Intelligence Subsystem is an augmentation of the current system to provide information and guidance as if a statistical consultant were present during an analysis. Basically we want to provide an expert subsystem that asks the right questions as we step through an analysis. In order to get insight -- more than answers -- asking the right questions is just as important as coming to the right conclusions. The expert system interacts with the analyst, setting the pace and posing questions along the way: have you checked this, have you checked that, what does such and such a plot look like? It will look like such and such so perhaps you should go in this direction, that direction, etc.

Display 34 shows some of the human-machine interaction problems that must be addressed in an expert system. If the user requests an operation like BOXPLOT, he should be able to see a one-line definition and the rationale for its use. In other words, if the expert system recommends a certain course of action, the analyst should be able to ask questions like, "What is the penalty if we don't follow this?."

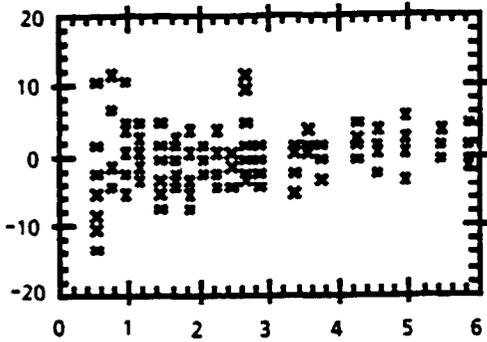


Fig. 2 Residuals of Exponential/Linear Fit.

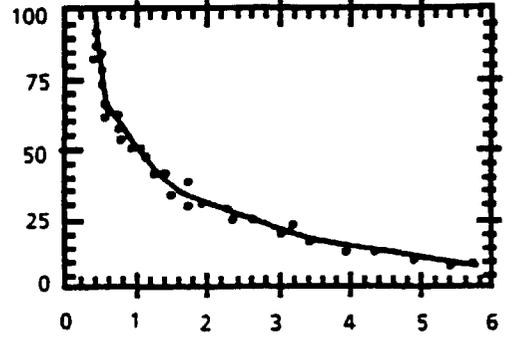
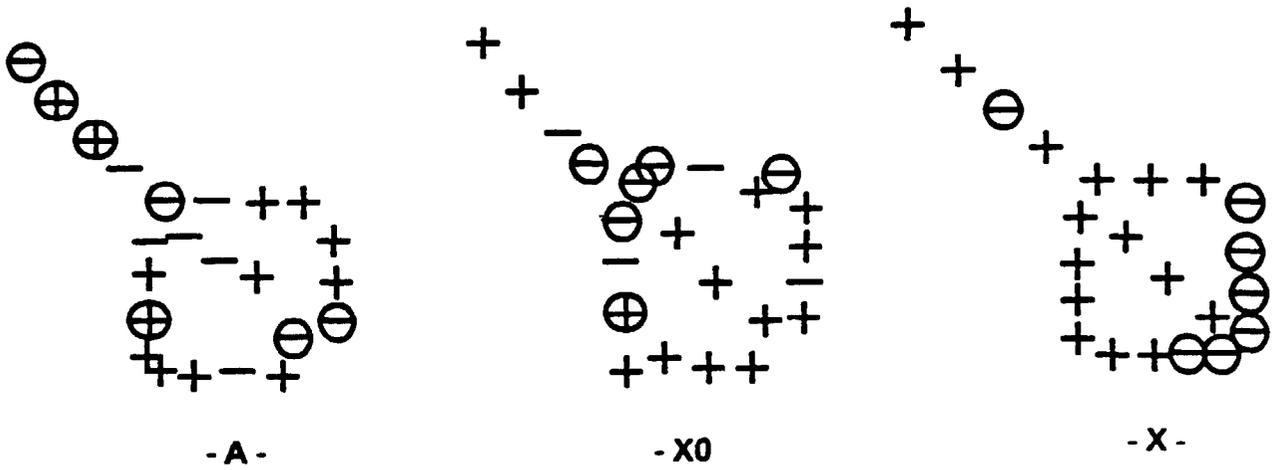


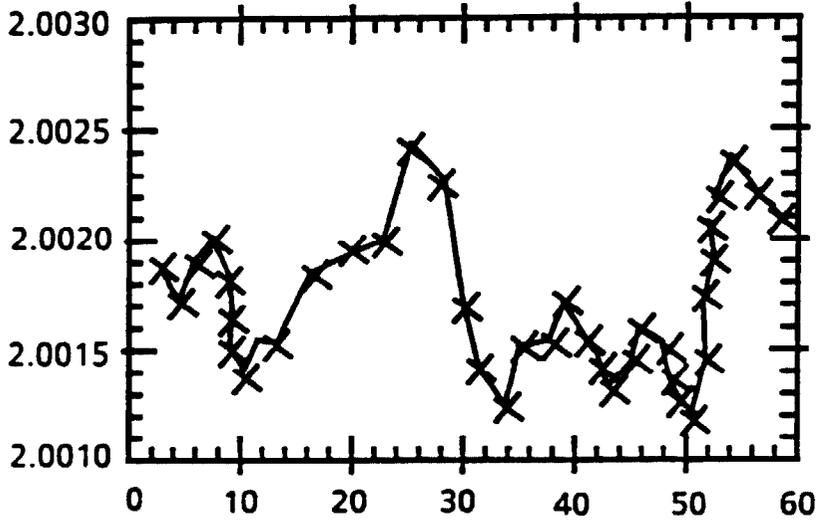
Fig. 3 Ultrasonic Response vs. Crack Length Data. After [2].

28MAR84 AWS 80:30 SRM 1761  
C 1930

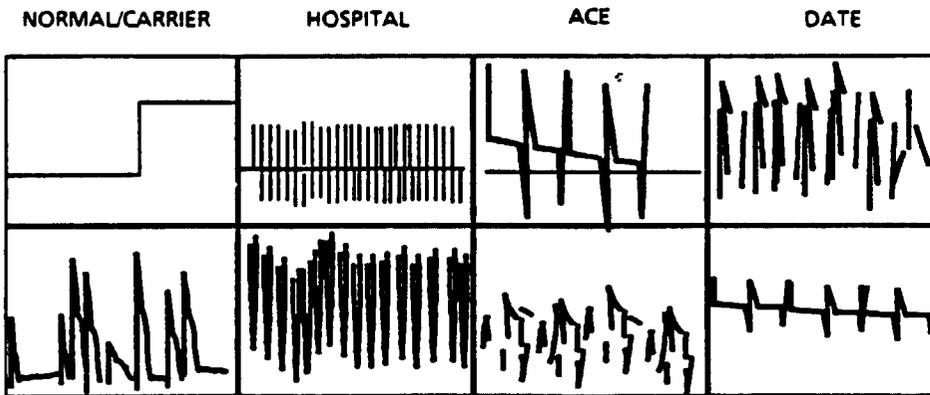


Display 32.

**QUESTION --DOES THE FIXED-LOCATION ASSUMPTION HOLD?  
TECHNIQUE--RUN SEQUENCE PLOT.**



Text Frame



**CONCLUSION: Shift in Location and Variation**

Display 33.

## HUMAN PROBLEMS (DESIGN GOALS)

**DEFINITIONS**

**RATIONALE**

**LINKING TOOLS**

**KEY TESTS**

**HYPOTHESES/CONCLUSIONS**

**VARYING EXPERIENCE**

*Display 34.*

The expert system should support the linking together of statistical analysis tools, often in unexpected ways. Data analysis is primarily sequential and interactive. We step through the data, step through the analysis; and at each step, the next step is dictated by what we have seen before that.

Scientists often deal with correlation plots to see if there is any correlation structure in the data. DATAPLOT supports a correlation-plot command and many other graphic commands and analysis. However, if someone asks for a correlation plot, the expert system should assist the analytic effort by carrying out appropriate statistical tests behind the scenes.

Another important but time-consuming aspect of data analysis (especially when you are writing research papers for the general science community) is the need to frame your hypothesis and conclusions in proper statistical terms. An expert system should support this formulation. We have found that to be very helpful to the average scientist and engineer. Every paper that goes out of NBS goes through our statistics review process to guarantee that hypotheses and conclusions have been properly stated in statistical terms.

The last aspect is the varying experience. Any expert system is going to have a problem dealing with different kinds of methods. No one expert system is going to be ideal because users have various degrees of experience. That's a very sticky problem. A tough problem. You don't want the expert system to be so simpleminded that an experienced analyst must go through 20 menus just to carry out a legal analysis. On the other hand, someone with limited experience needs the extensive guidance that 20 menus would give.

Display 35 shows the general content of the expert system component of DATAPLOT.

## SUBSYSTEM OUTPUT (THE EXPERT SYSTEM)

SEQUENCE OF MENUS

DATAPLOT COMMANDS

CAUTIONS/CONCERNS

MENU EXPLANATIONS

ADDITIONAL TESTS

RIGOROUS STATISTICAL CONCLUSIONS

*Display 35.*

*Sequence of Menus:* Each menu should have guidelines at the bottom of the menu explaining not only the current menu but offering suggestions as to which menus to select next for specific analyses and why. These suggestions can include specific DATAPLOT commands. Moreover, within the menu environment the cautions and concerns about the form of the analysis should be displayed clearly (e.g., a caution about the data not following a Normal distribution).

The user should always have access to *HELP* functions for each menu.. These menu explanations should include a description of where the particular menu fits within the entire collection of menus.

*Additional tests:* I mentioned the idea of performing statistical computations behind the scene. Although the analyst may be unaware of their specifics, he may want to make use of their results at a later time. The expert system is aware of this and can provide them.

DATAPLOT thus has 2 expert subsystems: A consultant-style expert system which offers expert guidance for thoroughly and rigorously carrying out a data analysis; and a data-interpretive expert system which chooses a test, applies the test to the data, interprets the output, and formulates a rigorous statistical conclusion (couched in proper statistical terminology).

The remaining displays provide some idea of the analyst's interaction with the expert system component of DATAPLOT.

As you can see the need for a great variety of interactions in a large expert system requires a lot of thought and a large comprehensive software system. If any of you want to see DATAPLOT in operation, we are out in Gaithersburg, and we will be glad to come out and demonstrate it locally.

## References

Filliben, J. J. and Fong, J. T. (1984), "DATAPLOT as an Expert System for Data Analysis," available from American Society of Mechanical Engineers, June, 1984.

Hahn, Gerald J. (1985), "More Intelligent Statistical Software and Statistical Expert Systems: Future Directions," The American Statistician, February, 1985.

### EDITING AND IMPUTATION

Brian Greenberg, Bureau of the Census

In talking today about an application of expert-system methods to data editing and imputation, it will be the first time that I use the words "expert system" in describing the edit and imputation program we have developed -- SPEER (Structured Program for Economic Editing and Referrals). In the past, the focus was more on describing the underlying methodology and discussing what the edit and imputation system could do for users.

While preparing notes for this talk, I found that the emphasis was less on SPEER itself and more on editing and imputation as an expert system in principle. When work started on our project to develop edit and imputation software we had no intention of building an expert system. The goal was to develop techniques that corrected survey and census response data and imputed for missing values. Looking back, one can see that as work on this project proceeded, an expert system was evolving; and in the talk I will describe some of the steps in the development of this system.

The purpose of editing and imputation is two-fold. First, if a respondent form is received and some responses are blank (item non-response), one tries to fill in missing values in order to create a complete data record for tabulation purposes. In addition, one wants to detect erroneous responses and correct them. For example, a response may indicate a fifty-acre farm with five million bushels of wheat or a twelve-year-old grandfather. Such problems do occur in the response data; they can be data entry errors or errors at the source. Which field does one adjust and what value should replace those selected for change?

When confronted with large data sets such as one has in the Census Bureau and many other Federal agencies, an automated system is a necessity. For surveys dealing with similar types of data, one would like to have general programs to avoid continually having to reinvent the wheel. On the other hand, it is desirable that an edit and imputation program incorporate as much survey-specific information as is available, and one would like the survey-specific information to be exercised through a family of rules. In addition, one usually would like a mathematical model to ensure that rules are applied consistently and to assist in selecting from among rules. In particular, one wants to blend survey-specific information with mathematical procedures within a coherent framework. The expert-system model is a natural structure for this type of program.

## FUNCTIONS IN AN EDIT AND IMPUTATION SYSTEM

### EDIT CHECKING

#### ERROR LOCALIZATION

#### IMPUTATION

*Display 36.*

What are the functions in an editing and imputation system? (See Display 36.) The first is *edit checking*. Edits are rules that detect prohibited response combinations; and it is easy to check when an edit fails, that is, a prohibited combination is encountered. Given an edit-failing record, one endeavors to change as few responses as possible in order to make the remaining responses consistent. Determining fields to change is called *error localization*. Finally, one wants to *impute* in order to allocate values for non-response and replace responses deleted during the error localization process.

## DESIGNING AN EXPERT SYSTEM FOR EDIT AND IMPUTATION

### UNDERSTANDING THEORETICAL ASPECTS OF EDITING AND IMPUTATION

### UNDERSTANDING FACETS AND NATURE OF SUBJECT-MATTER EXPERTISE

*Display 37.*

In designing edit and imputation software along the lines of an expert system, each function that was described above should be structured in its own module (see Display 37). In a general system one wants to enter as parameters the information that will be requested of all users. Survey-specific information, particularly decision rules, can be entered in specified, well-defined places throughout the program. These rules will be different for each user. SPEER had been employed on six segments of the 1982 Economic Censuses. The *edit-checking* routine never changed from user to user, and the *area-localization* subroutine was always the same. The *imputation* rules, however, varied in each application. How does one impute? In general, one must rely on those with expertise about the particular survey vehicle. One works with the subject-matter specialists to elicit well-defined decision rules based on their knowledge and experience.

What does one have to do in designing edit and imputation programs along the lines of an expert system? First, one must understand the nature and the facets of the subject-matter expertise. What do the experts know? Their experience concerning the survey vehicle is extensive; it is often based on experience in the analysis of response forms and familiarity with respondents. They are knowledgeable about the survey target population, the survey form itself, and often the source of errors or non-response in data.

As a matter of fact, for some kinds of missing data, the survey specialist can tell you why it's missing. For example, it may be known by people working on a survey that when a certain field is blank, the respondent means zero -- they just routinely skip the question. Other blank fields will never be zero. The respondent either did not know the answer to that question or did not want to reveal it; and so that data field was left blank. Knowledge of this sort is certainly survey-specific. It cannot be gleaned through standard analysis of reported data, nor are there usually auxiliary data sets available to design models of "missingness." The subject-specialist, however, is a source of information that can be profitably utilized. Statistically-derived procedures (such as appropriate model-based imputation techniques) can be viewed and utilized as survey-specific decision rules.

In addition to subject-matter expertise, one must incorporate appropriate editing models. In SPEER, the error-localization process is basically a set-covering problem -- a mathematical model. One utilizes linear analysis and graph theory to select fields to delete on edit-failing records. Once these fields are deleted, the remaining fields will be mutually consistent; and then one can begin to impute. The process of imputation uses survey-specific rules provided by subject-matter experts. The knowledge base of decision rules can be organized within coherent imputation modules through which they can be applied. That is, the system goes back and forth between the subject-matter information and the mathematical model. Mathematical techniques and subject-based imputation rules are two components that one should have in an overall edit and imputation system.

Thinking of it that way, the mathematical procedure and the subject-matter rules can be treated as separate. One can extend the mathematical methods and revise the flow of the system as a whole, unencumbered by survey-specific considerations. The survey-specific rules can be examined in their own right, updated and revised as needed independent from the programs through which they are applied. On the other hand, the mathematical procedures and decision rules are integrated. The mathematical constructions provide a framework to assist in choosing the most appropriate decision rule and to ensure that the value imputed will pass all applicable edits. Thus, an expert system for imputation should do more than provide a vehicle for accessing expert rules. It should also provide a mathematical framework to help decide from among the rules, choosing only rules which are valid for the record under consideration.

#### SPEER

#### CONTINUOUS (ECONOMIC) DATA UNDER RATIO EDITS

(A(1), ..., A(N))

#### TYPICAL RATIO EDIT

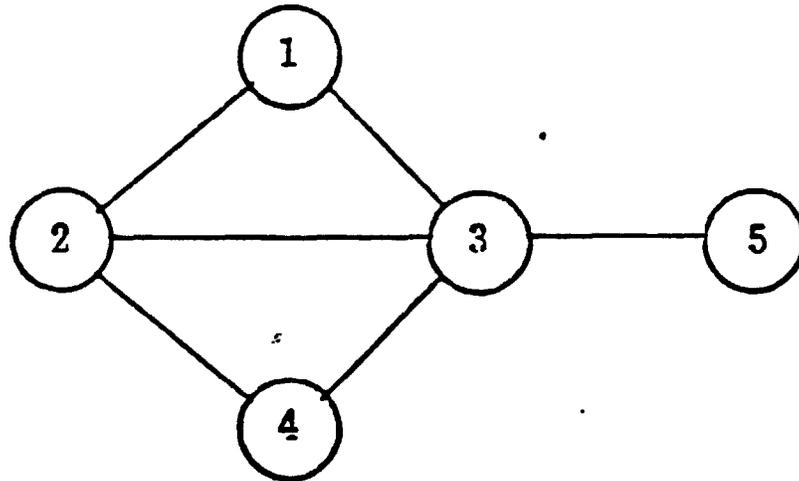
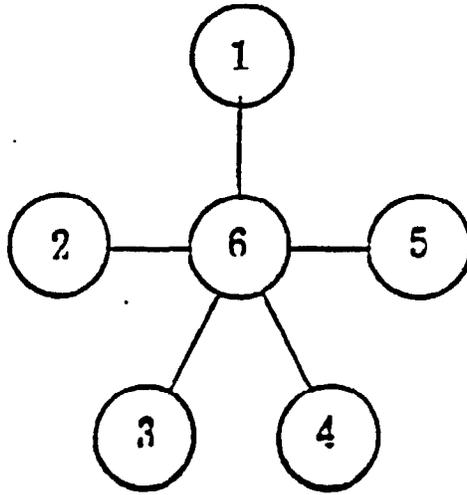
$$L(ij) \leq A(i)/A(j) \leq U(ij)$$

*Display 38.*

SPEER (Display 38) is an edit and imputation system designed along the lines of an expert system. SPEER was designed for economic data such as wages, assets, inventories, etc. The typical edit is a ratio of two fields, called a *ratio edit*. The total salaries paid to employees divided by the total number of employees should be within some reasonable range consistent with our knowledge of the industry and occupation. The amount of crop yield divided by the number of acres should be in a certain range. Ratio range checks are a very common edit in economic surveys. Given that a family of ratio edits is failed by a response record, one must select a set of fields to delete. We illustrate the workings of the error localization mechanism of SPEER on two samples below.

Let the circled numbers in Display 39 represent response fields and the edges in the graph represent *edit failures* between the adjoining fields. For example, the value in field 6 is inconsistent with fields 1 through 5 as determined by the collection of edit rules. If we delete the value in field 6 -- that is, remove node 6 from our graph -- all edges vanish. Thus, the remaining fields are mutually consistent because there are no edges connecting the corresponding nodes, hence there are no edit failures between them. That is, we can delete a single field to eliminate all edit failures. The lower graph is a little more complicated. One can see, however, that by deleting the values in fields 2 and 3, the remaining fields are mutually consistent. These simple examples capture the spirit of the error localization methods built into SPEER -- a little graph theory is used to find the minimal subset of field values to delete.

After error localization one has a collection of blank fields (some due to non-response, others because fields were deleted during the error localization process). The remaining fields are consistent with one another, and they must be consistent with values imputed. The program sets up a series of range specifications for a blank field taking into account the value for each valid field.



Display 39.

If  $A(n)$  is missing, and  $A(j)$  are consistent for all  $j \neq n$ ,

$$L(nj) \leq A(n) / A(j) \leq U(nj)$$

so

$$L(nj)A(j) \leq A(n) \leq U(nj)A(j)$$

Display 40.

Every valid imputation for a missing field (field  $n$  in Display 40) must lie in the overlap of the regions determined by each fixed field on the record in order to be consistent with every other field.

Once the *feasible region* for a missing field is computed the program reaches into the *imputation module* for the value to be imputed. The first applicable rule is selected and an imputation is derived based on this rule. If the derived value falls in the feasible region, it is accepted as a valid imputation. If not, the second rule is accessed, an imputation value is derived, etc. The value ultimately selected as the imputation will be chosen based on subject-matter based rules and will also be consistent with all other fields on the record under review, because it is forced to lie in the feasible region.

This may be a good time to provide an example of what a rule sequence might look like. Suppose one is to impute for a field such as Annual Payroll (APR) on an economic census or survey. For concreteness, let us couch our discussion in terms of the 1982 Economic Censuses. The first rule might be to derive an imputation based on the 1982 Administrative Data value for APR. If the value derived does not lie in the feasible region, one might try the 1981 Administrative Data value for APR. If this value is not suitable, we pass to a third option, etc., until a valid imputation is derived. Some imputation rules can be extremely field specific. For example, suppose some field is to be reported in tons. Assume that the feasible region allows valid responses to be between 500 and 1,000 tons and the value 1,800,000 was reported and deleted as an error. The applicable option might be to divide the reported value by 2,000 (subject-based information that respondents

sometimes report in pounds rather than tons). In this example, we would derive 900 tons and observing that this value is feasible, accept it as the valid imputation. A common error in reporting economic data is that respondents provide answers in units rather than in thousands as per instructions. For fields in which this error may occur, the first rule (when appropriate) is to divide the reported response by 1,000.

The editing and imputation for the 1982 Enterprise Summary Report and the 1982 Auxiliary Establishment Report (both portions of the 1982 Economic Censuses) was performed using SPEER. In addition, SPEER was used to process the Manufacturing, Retail, Wholesale, and Service segments of the 1982 Economic Censuses of Puerto Rico. In each of these applications, the edit checking, error localization routines, and basic system flow are the same. Each application, however, had its own family of decision rules for imputation. Each application employed different rules based on the survey-specific fields, relation between fields, and auxiliary information.

How does one implement an edit and imputation system based on expert-system principles? For a given application, start with the experts. One probes their expertise, elicits rules, and embeds those rules in the system components requiring them. Sample data is tested, performance evaluated, rules are refined as needed.

Editing of economic data records at the Census Bureau is a two-phase process. All records are run through an automated edit and imputation system in *batch mode*. Within the automated routines, selected records are targeted as *referral cases* and are directed for analyst review. An optimal strategy will include automated procedures to resolve the majority of cases and individual review for establishments needing special handling. Typical referral criteria are: (1) large change to reported data; (2) imputations for large establishments; and (3) highly atypical combination of responses. The analyst reviewing a response form is a subject-matter specialist, and the review is currently a pencil-and-paper process. After analyst adjustments are made to the results of automated processing on an establishment record, the revised form is once again processed through the automated system.

SPEER allows on-line, interactive processing of referral cases. Used in this mode, the system *converses* with an expert using it. The human expert can override the decision rules residing in the system and replace them based on his/her expertise and auxiliary information about the case under review. Using this system, the analyst requests a specific record and reviews the processing done by the automated system. The analyst has the original response form and hence access to information not incorporated into the rules. Based on this additional information and his/her own experience, an analyst may overrule the decision rules built into the automated system.

#### IMPUTATION OPTIONS FOR APR

- A. RANGE OF APR: (250,750)
  - B. CURRENT VALUE: 375
- OPTIONS
- 1. REPORTED VALUE: 82
  - 2. 1982 ADMINISTRATIVE DATA:
  - 3. 1981 ADMINISTRATIVE DATA BASED:
  - 4. 1977 CENSUS DATA BASED:
  - 5. IMPUTATION AND TOLERANCE:
  - 6. ANALYST SUPPLIED VALUE:

*Display 41.*

The display seen by the analyst looks something as in Display 41. Using Annual Payroll (APR), this display shows an acceptable range for APR from 250 to 750 (i.e., the feasible region). The current value is 375, which was derived by the automated system. The next value is the actual reported value of 82 followed by the reported 1982 Administrative Data and other candidate imputations based on 1981 Administrative Data, 1977 Economic Census Data, etc. The ordering above reflects the order in which the rule options are applied. By requiring that the range in which the imputed value must fall be consistent with all fields, plus a variety of options, the

analyst then has a significant amount of information at his/her disposal to assist in the decision-making process. If there is reason to believe that the most appropriate imputation value lies outside the feasible region (for example, because of explanatory notes on the form or through a call-back to the respondent), the analyst can select an imputed value outside the feasible region.

A revised imputation for field APR is decided, and the analyst enters it into the data record. This value is accepted by the program, and field APR is considered to be completed. Suppose there is a second field to be reviewed on this record (for example, Number of Employees (EMP)). Once again, the program displays on the terminal screen the feasible region for EMP, currently residing value, and candidate values for imputes derived according to each option, as it did for APR. Note, however, that each of these values is based, in part, on the new value of APR just entered by the analyst. As above, the analyst will determine an appropriate value for EMP, enter this value, and move on to the next field, if any. After all fields have been examined and adjusted if needed, the review is complete. The revised record will be consistent, and no further batch processing will be required.

The important observation from the perspective of an expert system, is that a true expert converses with the automated expert programs in order to augment the system expertise and override decision rules as needed. Initial testing has shown that analysts have found this system easy to use. It has the potential for making their decisions in the review of establishment records less tenuous than is currently the case. Because the individual review of establishment records is a time-consuming and costly process, one can anticipate savings of time and money in the use of an "expert-system aided," on-line, interactive review process. The on-line, interactive portion of this program has not yet been put to use for actual survey processing. We are actively working with potential users to incorporate this aspect of the program in future editing and imputation processing.

In summary, an edit and imputation system should blend statistical and subject-matter expertise in a coherent framework and integrate edit constraints with imputation strategy. We have described a structured system that attempts to meet these requirements and is sufficiently flexible to accommodate a variety of users. Development work continues on this system, enhancements are being made, and additional users are being identified. The references provide more information about some of the technical features of the SPEER system.

#### References

Greenberg, Brian (1981), "Developing an Edit System for Industry Statistics," *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, 11-16, Springer-Verlag, New York.

\_\_\_\_\_ (1982), "Using an Editing System to Develop Editing Specifications," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 366-371.

\_\_\_\_\_ and Surdi, Rita (1984), "A Flexible and Interactive Edit and Imputation System for Ratio Edits," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 421-426.

## DISCUSSION

Mark Winer, Office of Management and Budget

When Terry first mentioned the idea of having a panel on expert systems in a conference on Statistical Uses of Microcomputers in Federal Agencies, my question was "What do expert systems have to do with microcomputers and statistical systems?". I decided I would take a look at all the things I saw today to see how this session fits into the other sessions. The first thing we see is that you can use the IBM PC, or other personal computers, as a terminal to a mainframe. The mainframes have excellent software systems. That allows you to use both machines for the things those machines are best at. With the large amounts of data and large amounts of information you might need with an expert system, it is good to use a mainframe for most cases; but for the kind of processing and quick response you might want, it's nice to use downloaded results from an expert system on a personal computer.

The second reason this fit in is that, as I mentioned, the system developed by Brian Greenberg has just been adapted to personal computers. As memory capacity and storage capacity on personal computers increase, even the large systems like DATAPLOT could be extended to personal computers.

The third reason that this fit in is that every couple of years there is a real hot topic in the computer field. In 1982 and 1983 it was decision support systems. If we were having this workshop in 1982 or 1983, we would have undoubtedly had a panel on decision support systems; and since in 1984 and 1985 the hot topic is expert systems, we are having this workshop and it is incumbent upon you to have a panel on expert systems. This brings up the obvious question of whether an expert system really is something new or is it just something old, another big word that people use to bring out high-priced consultants to design your system.

I guess I will say that from what we have seen in these demonstrations today, expert systems are doing more than ordinary software systems. Ordinary software systems help the user do the things he normally does but make it possible for him to do those things faster and save him some of the tedious parts of the task. Both the systems we heard talked about today have the advantage of actually bringing additional knowledge to the user in that he can do what he wouldn't necessarily know how to do. Expert systems show you how to locate the error that is the easier error to change if you are trying to do efficient editing. You have subject-matter expertise that analysts couldn't produce on their own. This system does that for you. It uses the subject matter of the expert to figure out how that record can be changed with some help from the system.

The DATAPLOT system teaches you about the tests that are available to you as you use them. It suggests to you additional steps to do as you are doing things; so even if you are not an expert statistician, you can figure out the ways to proceed as you are working on a problem. So, as I say, expert systems provide something beyond what we ordinarily have in software systems. They are an extension of the existing packages rather than things that stand by themselves.

## QUESTIONS AND ANSWERS

Q1: You mentioned a new fad. Isn't some of this just like a sophisticated help facility?

A1: (Mr. Winer): Yes.

Q2: This is this year's new thing, but better help facilities have been a growing need since computers got started. I think expert systems are a logical outgrowth of that.

A2 (Mr. Lawton): I think expert systems are more than that, but help facilities are at least part of what we are dealing with here. I would say what makes the help facility more sophisticated is that they have some expertise built into them, that is based on knowledge of where in the program the user calls the help facility. So what you say is partly true, but I think there is some intelligence built into the back of the help facility in the expert system that wouldn't be there in a more conventional system or from just reading the help file.

A2 (Mr. Ireland): There probably are two other issues. First, the help system can be changed incrementally as you come to understand what kind of help you need. The idea of rules makes it easy to develop small help modules that are added to a system that already has a help facility. Second, for some of Brian's things, it isn't a user help facility, but a specification of how to handle a particular piece of data. So, the help facility might never be seen by analysts unless they ask to see it, but it would be used to make a proper kind of modification to the data.

A2 (Mr. Greenberg): Expert systems can be run in batch mode once expertise is built into it, and that bears on the use of the help facility in a batch mode.

Q3: I am curious about some of the details of the DATAPLOT and the editing and imputation system. Let me start with the editing system since that is fresh in your memory. How much memory on the IBM XT does your system take up?

A3 (Mr. Greenberg): I really don't know. I haven't been doing the actual transfer to the XT.

Q4: But you could fit it into 640K?

A4 (Mr. Greenberg): Yes, plenty of data is on one floppy disk.

Q5: You said it was easy to learn to use. What did you say--a half day? How long?

A5 (Mr. Greenberg): I would say a half day working with somebody like myself or someone familiar with it.

Q6: We do surveys a lot and they are typically tedious -- we have people coming in to do error checking and editing in a rather primitive way, so I think your product would be very useful to us and a lot of other people. What would be a good way for us to learn more about it?

A6 (Mr. Greenberg): Drop me a line or give me a call.

Q7: The degree of statistical information you obviously have in your head goes well beyond that of everyone in our office. The only fear I would have would be whether those of us who have a much lower level of statistical knowledge could still make use of DATAPLOT. What do you think about that?

A7 (Mr. Filliben): That is a general problem, and one of the displays dealt with varying experience. This addresses the point of whether this is an extended help facility. We tried to make sure that the menus that came up would be a part of the education process too--a tutorial, if you will. We have had people use this expert system who have very limited statistical background, and they have come out with good results. It's a matter of learning, and I think the expert systems are at the point now where it's nice to have a machine that has an expert system, but it's also nice to have some statisticians and other consultants around who can augment them. One thing we did not mention was references. Where does someone go if he really wants to read up on graphical or residual analysis, for example? That is one command as far as DATAPLOT is concerned. There should be a reference command. It's not in there yet, but there is a body of literature that's out there that has a lot of details. If people want to go in and fill up their own base knowledge, they should have access to this base. It is very much, as you say, an extended help. There are lots of different ways these various systems can be of help because there are a lot of different ways we can have deficiencies in our own knowledge.

Q8: What kind of mainframe are they working with?

A8 (Mr. Filliben): All the major mainframes. UNIVACs. The most popular one, and in fact the default machine, is the VAX 11/780. IBM/PC'S. The Pentagon has it on a Honeywell Multics system. PERKIN-ELMERS. PRIMES. The only machine we had difficulty putting it up on was the CYBER machine. That problem will disappear because we are getting a CYBER machine and we will be forced to address that problem. They have a hardware restriction on memory. In UNIVAC you run into an overlay problem. In terms of whether it would download to a PC or could be put up on a PC, you would need about a half a megabyte of memory. Small machines--micros--are expanding to the point where it's a real possibility to put DATAPLOT on a PC.

Q9: You say that NTIS sells this?

A9 (Mr. Filliben): National Technical Information Service sells it for \$1200 -- a one-time-only fee. You get the source code. The source code on the file is 12 megabytes, so you have to have somewhere where you can put it.

Q10: Did you write this yourself?

A10 (Mr. Filliben): Yes.

Q11: How long did it take?

All (Mr. Filliben): We started back in 1972 with a software system called DATAPAK which is free from NBS. That sort of got us into the problem. By 1975 it became clear that interactive systems were becoming more important. By 1977 we had the first DATAPLOT running, and things have essentially been the same since then. We augmented it to include the expert system.

Mr. Winer: Perhaps less a question than a comment. At the end of the last session, I asked the panelists how their decentralized and spreadsheet-type statistical systems insure or assure data integrity and adherence to the statistical standards. Here I think we have had two presentations in which one could in a sense say "Hey, that answers the question!". If people start using systems like Brian's, they will have more data integrity; and if one starts using systems more like Jim's, one could have more adherence to agencies' present standards.

I would like to take this opportunity to thank Terry Ireland who chaired this session, but who is also the Chairman of the Subcommittee of the Federal Committee on Statistical Methodology who organized this entire workshop, including this session. We thank him and thank you all for coming.

## Appendix

### Announcement of Workshop on Statistical Uses of Microcomputers in Federal Agencies

The Subcommittee on Statistical Uses of Microcomputers in Federal Agencies of the Federal Committee on Statistical Methodology is sponsoring a one-day workshop on April 24, 1985, to discuss with other Federal employees selected topics on statistical uses of microcomputers. The workshop will be held at the IRS Auditorium, 1111 Constitution Avenue, N.W., 7th floor, from 9:15 a.m. to 4:30 p.m.

The agenda and speakers are as follows:

9:15 a.m. WELCOME AND INTRODUCTION

Chair: Maria Gonzalez, Office of Management and Budget

Arrangements: Linda Taylor, Internal Revenue Service

9:20 a.m. PLANNING

Chair and Discussant: Larry Cox, Bureau of the Census

Rapporteur: Fred Cavanaugh, Bureau of the Census

Microcomputer technology has much to offer statistics, and many statisticians have become microcomputer users at work and at home. This technology and the keen interest of statisticians in it provide statistical agencies with many opportunities, each bringing with it responsibilities for planning, implementation and evaluation: If every statistician (programmer/secretary) in an agency wants a microcomputer, who should have them? For what purposes can/should microcomputers be used? In what configuration? At what cost (overall/per user)? How will this technology coexist with central ADP services? What policy decisions need to be made -- when -- by whom?

In this session on planning, we will explore such questions through discussion, focusing on three different and successful approaches to these problems -- those adopted by the Census Bureau, the National Security Agency and the Bureau of Labor Statistics.

Speakers: Ronald R. Swank, Bureau of the Census

Kathy Schnaubelt, National Security Agency

Peter Stevens, Bureau of Labor Statistics

\* \* \* DISCUSSION \* \* \*

10:45 a.m. \* \* \* COFFEE BREAK \* \* \*

11:00 a.m.

ELECTRONIC DATA DISSEMINATION

Chair: Ken Berkman, Bureau of Economic Analysis

Rapporteur: Jay Casselberry, Energy Information Agency

This session is a panel discussion of the different approaches to electronic data dissemination by various Federal agencies. Different approaches will be described with particular emphasis on the factors determining an agency's approach to data dissemination and the problems encountered in their implementation. The experience gained by these agencies will be presented: National Technical Information Service (NTIS) distribution of microcomputer floppy disks; Census' CENDATA system; and the U. S. Department of Agriculture's current development of an on-line system.

Speakers: Stuart Weisman, National Technical Information Service

Barbara Aldrich, Bureau of the Census

Roxanne Williams, Department of Agriculture

\* \* \* DISCUSSION \* \* \*

12:30 p.m.

\* \* \* LUNCH \* \* \*

1:15 p.m.

APPLICATIONS OF MICROCOMPUTERS

Chair: Ron Steele, Department of Agriculture

Rapporteur: Tom Nagle, Internal Revenue Service

This session is a panel discussion of statistical applications of microcomputers. The utility and weaknesses of applications software and operating systems will be discussed. Some examples involve interfacing mainframe and microcomputers. Issues to be addressed include an assessment of the utility of microcomputers at present, the future utility in light of new hardware and software technologies, and considerations regarding data integrity, security and accessibility.

Speakers: Linda Atkinson, Department of Agriculture

Gary Nelson, Department of Agriculture

Rick Hayes, Internal Revenue Service

Brian Carney, Department of Agriculture

Paul Dobbins, Department of the Treasury

Dick Shively, Department of Agriculture

\* \* \* DISCUSSION \* \* \*

2:45 p.m.

\* \* \* COFFEE BREAK \* \* \*

3:00 p.m.

EXPERT SYSTEMS

Chair: Terry Ireland, National Security Agency

Discussant: Mark Winer, Office of Management and Budget

Rapporteur: Norman Glick, National Security Agency

Recently, the idea of incorporating techniques used by professional experts into software has become popular. This session will introduce the basis for expert-system methodology and give several practical examples of expert systems with statistical applications that are currently in use.

Speakers: George Lawton, Army Research Institute

James Filliben, National Bureau of Standards

Brian Greenberg, Bureau of the Census

\* \* \* DISCUSSION \* \* \*

4:30 p.m.

\* \* \* ADJOURN \* \* \*



**Reports Available in the  
Statistical Policy  
Working Paper Series**

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (Available through NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (Available through NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (Available through NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (Available through NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (Available through NTIS Document Sales, PB86-214285/AS)
7. An Interagency Review of Time-Series Revision Policies (Available through NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (Available through NTIS Document Sales, PB86-230570/AS)
9. Contracting for Surveys (Available through NTIS Document Sales, PB83-233148)
10. Approaches to Developing Questionnaires (Available through NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (Available through NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (Available through NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (Available through NTIS Document Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (Available through NTIS Document Sales, PB87-166393)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650